Bound
Periodical 1262440

Kansas City
Public Library

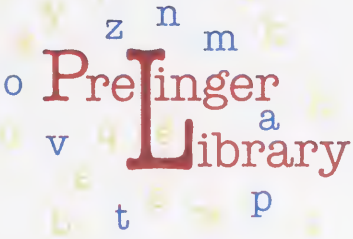
THE SCHOOL DISTRICT OF KANSAS CITY, MISSOURI

This Volume is for

REFERENCE USE ONLY

PUBLIC LIBRARY
KANSAS CITY
MO

From the collection of the



San Francisco, California
2008

PUBLIC LIBRARY
KANSAS CITY
MO

PUBLIC LIBRARY
KANSAS CITY
MO

PUBLIC LIBRARY
KANSAS CITY
MO

THE BELL SYSTEM TECHNICAL JOURNAL

A JOURNAL DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

EDITORS

R. W. King J. O. Perrine

EDITORIAL BOARD

W. H. Harrison
O. B. Blackwell
H. S. Osborne
J. J. Pilliod
O. E. Buckley
M. J. Kelly
A. B. Clark
F. J. Feely

TABLE OF CONTENTS AND INDEX

VOLUME XXVI

1947

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PUBLIC LIBRARY
KANSAS CITY
MO

PRINTED IN U. S. A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXVI 1947

Table of Contents

JANUARY, 1947

Development of Silicon Crystal Rectifiers for Microwave Radar Receivers—*J. H. Scaff and R. S. Ohl* 1
End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator—*J. P. Kinzer and I. G. Wilson* 31
First and Second Order Equations for Piezoelectric Crystals Expressed in Tensor Form—*W. P. Mason* 80
The Biased Ideal Rectifier—*W. R. Bennett* 139
Properties and Uses of Thermistors—Thermally Sensitive Resistors—*J. A. Becker, C. B. Green and G. L. Pearson* 170

APRIL, 1947

Radar Antennas—*H. T. Friis and W. D. Lewis* 219
Probability Functions for the Modulus and Angle of the Normal Complex Variate—*Ray S. Hoyt* 318
Spectrum Analysis of Pulse Modulated Waves—*J. C. Lozier* 360

JULY, 1947

Telephony by Pulse Code Modulation—*W. M. Goodall* 395
Some Results on Cylindrical Cavity Resonators—*J. P. Kinzer and I. G. Wilson* 410
Precision Measurement of Impedance Mismatches in Waveguide—*Allen F. Pomeroy* 446
Reflex Oscillators—*J. R. Pierce and W. G. Shepherd* 460

1262440 MAR 9 1948

OCTOBER, 1947

The Radar Receiver—*L. W. Morrison, Jr.* 693
High-Vacuum Oxide-Cathode Pulse Modulator Tubes—*C. E. Fay* 818
Polyrod Antennas—*G. E. Mueller and W. A. Tyrrell* 837
Targets for Microwave Radar Navigation—*Sloan D. Robertson* 852
Tables of Phase Associated with a Semi-Infinite Unit Slope of Attenuation—*D. E. Thomas* 870

Index to Volume XXVI

A

Analysis, Spectrum, of Pulse Modulated Waves, *J. C. Lozier*, page 360.
Antennas, Polyrod, *G. E. Mueller and W. A. Tyrrell*, page 837.
Antennas, Radar, *H. T. Friis and W. D. Lewis*, page 219.
Attenuation, Tables of Phase Associated with a Semi-Infinite Unit Slope of, *D. E. Thomas*, page 870.

B

Becker, J. A., C. B. Green and G. L. Pearson, Properties and Uses of Thermistors—Thermally Sensitive Resistors, page 170.
Bennett, W. R., The Biased Ideal Rectifier, page 139.

C

Cavity Resonator, Circular Cylinder, End Plate and Side Wall Currents in, *J. P. Kinzer and I. G. Wilson*, page 31.
Cavity Resonators, Cylindrical, Some Results on, *J. P. Kinzer and I. G. Wilson*, page 410.
Code Modulation, Pulse, Telephony by, *W. M. Goodall*, page 395.
Crystal, Silicon, Rectifiers for Microwave Radar Receivers, Development of, *J. H. Scaff and R. S. Ohl*, page 1.
Crystals, Piezoelectric, Expressed in Tensor Form, First and Second Order Equations for, *W. P. Mason*, page 80.

F

Fay, C. E., High-Vacuum Oxide-Cathode Pulse Modulator Tubes, page 818.
Friis, H. T. and W. D. Lewis, Radar Antennas, page 219.

G

Goodall, W. M., Telephony by Pulse Code Modulation, page 395.
Green, C. B., G. L. Pearson and J. A. Becker, Properties and Uses of Thermistors—Thermally Sensitive Resistors, page 170.

H

Hoyt, Ray S., Probability Functions for the Modulus and Angle of the Normal Complex Variate, page 318.

I

Impedance Mismatches in Waveguide, Precision Measurement of, *Allen F. Pomeroy*, page 446.

K

Kinzer, J. P. and I. G. Wilson, End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator, page 31.
Kinzer, J. P. and I. G. Wilson, Some Results on Cylindrical Cavity Resonators, page 410.

L

Lewis, W. D. and H. T. Friis, Radar Antennas, page 219.
Lozier, J. C., Spectrum Analysis of Pulse Modulated Waves, page 360.

M

Mason, W. P., First and Second Order Equations for Piezoelectric Crystals Expressed in Tensor Form, page 80.

Microwave Radar Navigation, Targets for, *Sloan D. Robertson*, page 852.
Microwave Radar Receivers, Development of Silicon Crystal Rectifiers for, *J. H. Scaff and R. S. Ohl*, page 1.
Mismatches, Impedance, in Waveguide, Precision Measurement of, *Allen F. Pomeroy*, page 446.
Modulated Waves, Pulse, Spectrum Analysis of, *J. C. Lozier*, page 360.
Modulation, Pulse Code, Telephony by, *W. M. Goodall*, page 395.
Modulator Tubes, High-Vacuum Oxide-Cathode Pulse, *C. E. Fay*, page 818.
Morrison, Jr., L. W., The Radar Receiver, page 693.
Mueller, G. E. and W. A. Tyrrell, Polyrod Antennas, page 837.

N

Navigation, Microwave Radar, Targets for, *Sloan D. Robertson*, page 852.

O

Ohl, R. S. and J. H. Scaff, Development of Silicon Crystal Rectifiers for Microwave Radar Receivers, page 1.
Oscillators, Reflex, *J. R. Pierce and W. G. Shepherd*, page 460.

P

Pearson, G. L., J. A. Becker and C. B. Green, Properties and Uses of Thermistors—Thermally Sensitive Resistors, page 170.
Phase, Tables of, Associated with a Semi-Infinite Unit Slope of Attenuation, *D. E. Thomas*, page 870.
Pierce, J. R. and W. G. Shepherd, Reflex Oscillators, page 460.
Piezoelectric Crystals Expressed in Tensor Form, First and Second Order Equations for, *W. P. Mason*, page 80.
Polyrod Antennas, *G. E. Mueller and W. A. Tyrrell*, page 837.
Pomeroy, Allen F., Precision Measurement of Impedance Mismatches in Waveguide, page 446.
Probability Functions for the Modulus and Angle of the Normal Complex Variate, *Ray S. Hoyt*, page 318.
Pulse Code Modulation, Telephony by, *W. M. Goodall*, page 395.
Pulse Modulated Waves, Spectrum Analysis of, *J. C. Lozier*, page 360.
Pulse Modulator Tubes, High-Vacuum Oxide-Cathode, *C. E. Fay*, page 818.

R

Radar: High-Vacuum Oxide-Cathode Pulse Modulator Tubes, *C. E. Fay*, page 818.
Radar: End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator, *J. P. Kinzer and I. G. Wilson*, page 31.
Radar: Some Results on Cylindrical Cavity Resonators, *J. P. Kinzer and I. G. Wilson*, page 410.
Radar: Polyrod Antennas, *G. E. Mueller and W. A. Tyrrell*, page 837.
Radar: Reflex Oscillators, *J. R. Pierce and W. G. Shepherd*, page 460.
Radar Antennas, *H. T. Friis and W. D. Lewis*, page 219.
Radar Navigation, Microwave, Targets for, *Sloan D. Robertson*, page 852.
Radar Receiver, The, *L. W. Morrison, Jr.*, page 693.
Radar Receivers, Microwave, Development of Silicon Crystal Rectifiers for, *J. H. Scaff and R. S. Ohl*, page 1.
Receiver, Radar, The, *L. W. Morrison, Jr.*, page 693.
Receivers, Microwave Radar, Development of Silicon Crystal Rectifiers for, *J. H. Scaff and R. S. Ohl*, page 1.
Rectifier, Biased Ideal, The, *W. R. Bennett*, page 139.
Rectifiers, Silicon Crystal, for Microwave Radar Receivers, Development of, *J. H. Scaff and R. S. Ohl*, page 1.
Reflex Oscillators, *J. R. Pierce and W. G. Shepherd*, page 460.
Resistors, Thermally Sensitive—Properties and Uses of Thermistors, *J. A. Becker, C. B. Green and G. L. Pearson*, page 170.
Resonator, Circular Cylinder Cavity, End Plate and Side Wall Currents in, *J. P. Kinzer and I. G. Wilson*, page 31.
Resonators, Cylindrical Cavity, Some Results on, *J. P. Kinzer and I. G. Wilson*, page 410.
Robertson, Sloan D., Targets for Microwave Radar Navigation, page 852.

S

Scaff, J. H. and R. S. Ohl, Development of Silicon Crystal Rectifiers for Microwave Radar Receivers, page 1.
Shepherd, W. G. and J. R. Pierce, Reflex Oscillators, page 460.
Silicon Crystal Rectifiers for Microwave Radar Receivers, Development of, *J. H. Scaff and R. S. Ohl*, page 1.
Spectrum Analysis of Pulse Modulated Waves, *J. C. Lozier*, page 360.

T

Tensor Form, First and Second Order Equations for Piezoelectric Crystals Expressed in, *W. P. Mason*, page 80.
Thermistors, Properties and Uses of—Thermally Sensitive Resistors, *J. A. Becker, C. B. Green and G. L. Pearson*, page 170.
Thomas, D. E., Tables of Phase Associated with a Semi-Infinite Unit Slope of Attenuation, page 870.
Tyrrell, W. A. and G. E. Mueller, Polyrod Antennas, page 837.

V

Vacuum, High-, Oxide-Cathode Pulse Modulator Tubes, *C. E. Fay*, page 818.

W

Waveguide, Precision Measurement of Impedance Mismatches in, *Allen F. Pomeroy*, page 446.
Wilson, I. G. and J. P. Kinzer, End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator, page 31.
Wilson, I. G. and J. P. Kinzer, Some Results on Cylindrical Cavity Resonators, page 410.

V. 26

VOLUME XXVI JANUARY, 1947 NO. 1

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Public Library
Kansas City, Mo.

Development of Silicon Crystal Rectifiers for Microwave Radar Receivers.....*J. H. Scaff and R. S. Ohl* 1

End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator.....*J. P. Kinzer and I. G. Wilson* 31

First and Second Order Equations for Piezoelectric Crystals Expressed in Tensor Form.....*W. P. Mason* 80

The Biased Ideal Rectifier.....*W. R. Bennett* 139

Properties and Uses of Thermistors—Thermally Sensitive Resistors..*J. A. Becker, C. B. Green and G. L. Pearson* 170

Abstracts of Technical Articles by Bell System Authors.. 213

Contributors to This Issue..... 217

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50¢ per copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORS

R. W. King J. O. Perrine

EDITORIAL BOARD

W. H. Harrison O. E. Buckley
O. B. Blackwell M. J. Kelly
H. S. Osborne A. B. Clark
J. J. Pilliod S. Bracken

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are 50 cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1947
American Telephone and Telegraph Company

PRINTED IN U. S A

CORRECTION FOR ISSUE OF OCTOBER, 1946

In the article *SPARK GAP SWITCHES FOR RADAR*, lines 2–14 inclusive on page 593 should have appeared between lines 10 and 11 on page 588.

The Bell System Technical Journal

Vol. XXVI *January, 1947* *No. 1*

Development of Silicon Crystal Rectifiers for Microwave Radar Receivers

By J. H. SCAFF and R. S. OHL

Introduction

TO THOSE not familiar with the design of microwave radars the extensive war use of recently developed crystal rectifiers[1] in radar receiver frequency converters may be surprising. In the renaissance of this once familiar component of early radio receiving sets there have been developments in materials, processes, and structural design leading to vastly improved converters through greater sensitivity, stability, and ruggedness of the rectifier unit. As a result of these developments a series of crystal rectifiers was engineered for production in large quantities to the exacting electrical specifications demanded by advanced microwave techniques and to the mechanical requirements demanded of combat equipment.

The work on crystal rectifiers at Bell Telephone Laboratories during the war was a part of an extensive cooperative research and development program on microwave weapons. The Office of Scientific Research and Development, through the Radiation Laboratory at the Massachusetts Institute of Technology, served as the coordinating agency for work conducted at various university, government, and industrial laboratories in this country and as a liaison agency with British and other Allied organizations. However, prior to the inception of this cooperative program, basic studies on the use of crystal rectifiers had been conducted in Bell Telephone Laboratories. The series of crystal rectifiers now available may thus be considered to be the outgrowth of work conducted in three distinct periods. First, in the interval from 1934 to the end of 1940, devices incorporating point contact rectifiers came into general use in the researches in ultra-high-frequency and microwave communications techniques then under way at the Holmdel Radio Laboratories of Bell Telephone Laboratories.

[1] A crystal rectifier is an assymmetrical, non-linear circuit element in which the seat of rectification is immediately underneath a point contact applied to the surface of a semiconductor. This element is frequently called "point contact rectifier" and "crystal detector" also. In this paper these terms are considered to be synonymous.

At that time the improvement in sensitivity of microwave receivers employing crystal rectifiers in the frequency converters was clearly recognized, as were the advantages of rectifiers using silicon rather than certain well known minerals as the semi-conductor. In the second period, from 1941 to 1942, the advent of important war uses for microwave devices stimulated increased activity in both research and development. During these years the pattern for the interchange of technical information on microwave devices through government sponsored channels was established and was continued through the entire period of the war. With the extensive interchange of information, considerable international standardization was achieved. In view of the urgent equipment needs of the Armed Services emphasis was placed on an early standardization of designs for production. This resulted in the first of the modern series of rectifiers, namely, the ceramic cartridge design later coded through the Radio Manufacturers Association as type 1N21. In the third period, from 1942 to the present time, process and design advances accruing from intensive research and development made possible the coding and manufacture of an extensive series of rectifiers all markedly superior to the original 1N21 unit.

It is the purpose of this paper to review the work done in Bell Telephone Laboratories on silicon point contact rectifiers during the three periods mentioned above, and to discuss briefly typical properties of the rectifiers, several of the more important applications and the production history.

Crystal Rectifiers in the Early Microwave Research

The technical need for the modern crystal rectifier arose in research on ultra-high frequency communications techniques. Here as the frontier of the technically useful portion of the radio spectrum was steadily advanced into the microwave region, certain limitations in conventional vacuum tube detectors assumed increasing importance. Fundamentally, these limitations resulted from the large interelectrode capacitance and the finite time of transit of electrons between cathode and anode within the tubes. At the microwave frequencies (3000 megacycles and higher), they became of first importance. As transit time effects are virtually absent in point contact rectifiers, and since the capacitance is minute, it was logical that the utility of these devices should again be explored for laboratory use.

The design of the point contact rectifiers used in these researches was dictated largely, of course, by the needs of the laboratory. Frequently the rectifier housing formed an integral part of the electrical circuit design while other structures took the form of a replaceable resistor-like cartridge. A variety of structures, including the modern types, arranged in chronological sequence, are shown in the photograph, Fig. 1. In general, the

principal requirements of the rectifiers for laboratory use were that the units be sensitive, stable chemically, mechanically, and electrically, and

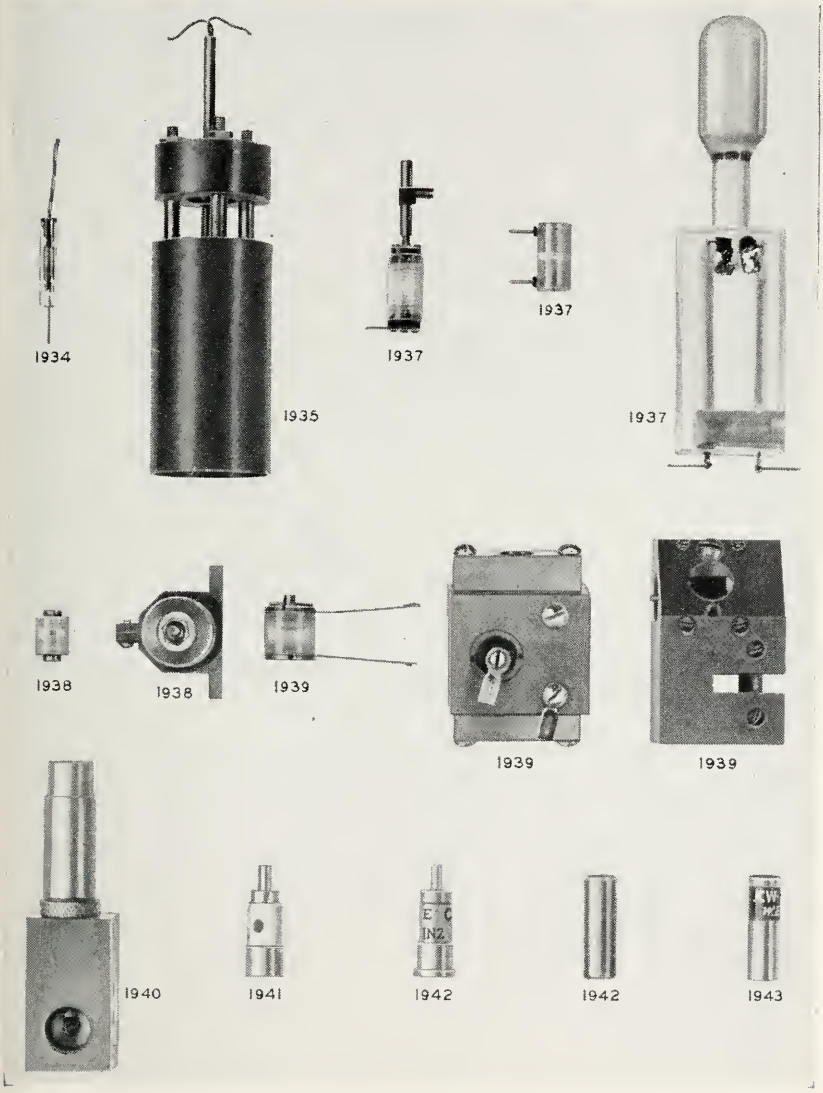


Fig. 1—Point contact rectifier structures. 1934–1943. Approximately $\frac{3}{5}$ actual size.

that they be easily adjusted. Considering the known vagaries of the device's historical counterpart, it was considered prudent to provide in the structures means by which the unit could be readjusted as frequently as might prove necessary or desirable.

As the properties of various semi-conductors were known to vary widely, an essential part of the early work was a survey of the properties of a number of minerals and metalloids potentially useful as rectifier materials. There were examined and tested approximately 100 materials, including zincite, molybdenite, galena, iron pyrites, silicon carbide, and silicon. Of the materials investigated most were found to be unsuitable for one reason or another, and iron pyrites and silicon were selected as having the best overall characteristics. The subsequent studies were then directed toward improving the rectifying material, the rectifying surface, the point contact and the mounting structure.

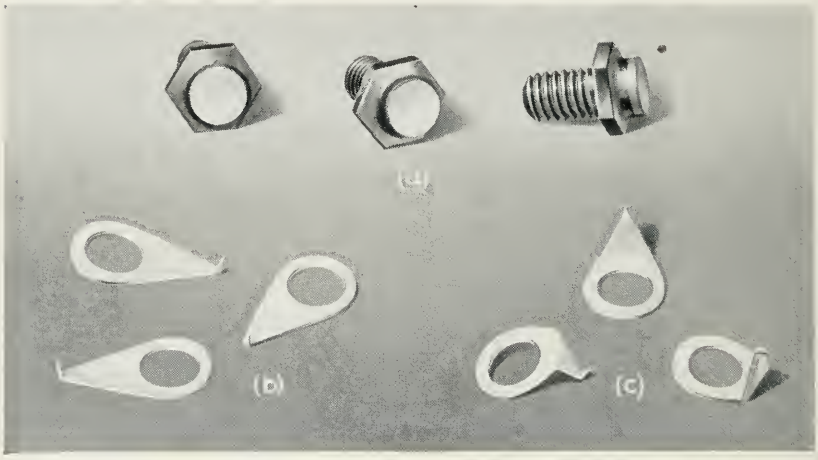


Fig. 2—Rectifier inserts and contact points for use in early 3000 megacycle converters. Overall length of insert $\frac{3}{16}$-inch approximately.

For use at frequencies in the region of 3000 megacycles standard demountable elements, consisting of rectifier "inserts" and contact points, were developed for use in various housings or mounting blocks, depending upon the particular circuit requirements. The rectifier "inserts" consisted of small wafers of iron pyrite or silicon, soldered to hexagonal brass studs as shown in Fig. 2a. In these devices the surface of the semi-conductor was prepared by grinding, polishing, and etching to develop good rectification characteristics. Our knowledge of the metallurgy of silicon had advanced by this time to the stage where a uniformly active rectifier surface could be produced and searching for active spots was not necessary. Furthermore, it was possible to prepare inserts of a positive or negative variety, signifying that the easy direction of current flow was obtained with the silicon positive with respect to the point or vice versa. Owing to a greater nonlinearity of the current voltage characteristic, the n-type or negative

insert tended to give better performance as microwave converters while the p-type, or positive insert, because of greater sensitivity at low voltages, proved to be more useful in test equipment such as resonance indicators in frequency meters. In certain instances also, it was advantageous for the designer to be able to choose the polarity best suited to his circuit design. In contrast, however, to the striking uniformity obtained with the silicon processed in the laboratory, the pyrite inserts were very non-uniform. Active rectification spots on these natural mineral specimens could be found only by tediously searching the surface of the specimen. Moreover, rectifiers employing the pyrite inserts showed a greater variation in properties with frequency than those in which silicon was used.

In addition to providing a satisfactory semi-conductor, it was necessary also to develop suitable materials for use as point contacts. For this use metals were required which had satisfactory rectification characteristics with respect to silicon or pyrites and sufficient hardness so that excessive contact areas were not obtained at the contact pressures employed in the rectifier assembly. The metals finally chosen were a platinum-iridium alloy and tungsten, which in some cases was coated with a gold alloy. These were employed in the form of a fine wire spot welded to a suitable spring member. The spring members themselves were usually of a wedge shaped cantilever design and were made from coin silver to facilitate electrical connection to the spring. Several contact springs of two typical designs are shown in the photograph, Figs. 2b and 2c.

A typical mounting block arranged for use with the inserts and points is shown in Fig. 1 (1940) and in Fig. 3. This block was so constructed that it could be inserted in a 70 ohm coaxial line without introducing serious discontinuities in the line. The contact point of the rectifier was assembled in the block to be electrically connected to the central conductor of the coaxial radio frequency input fitting, while the crystal insert screwed into a tapered brass pin electrically connected to the central conductor of the coaxial intermediate frequency and d-c output fitting. The tapered pin fitted tightly into a tapered hole in a supporting brass cylinder, but was insulated from the cylinder by a few turns of polystyrene tape several thousandths of an inch thick. This central pin was thus one terminal of a coaxial high-frequency by-pass condenser. The capacitance of this condenser depended upon the general nature of the circuits in which the block was to be used, and was generally about 15 mmfs. The arrangement of the point, the crystal insert and their respective supporting members was such that the point contact could be made to engage the surface of the silicon at any spot and at the contact pressure desired and thereafter be clamped firmly in a fixed position by set screws. Typical direct current characteristics of the positive and negative silicon inserts and of pyrite inserts assembled and adjusted in this mounting block are shown in Fig. 4.

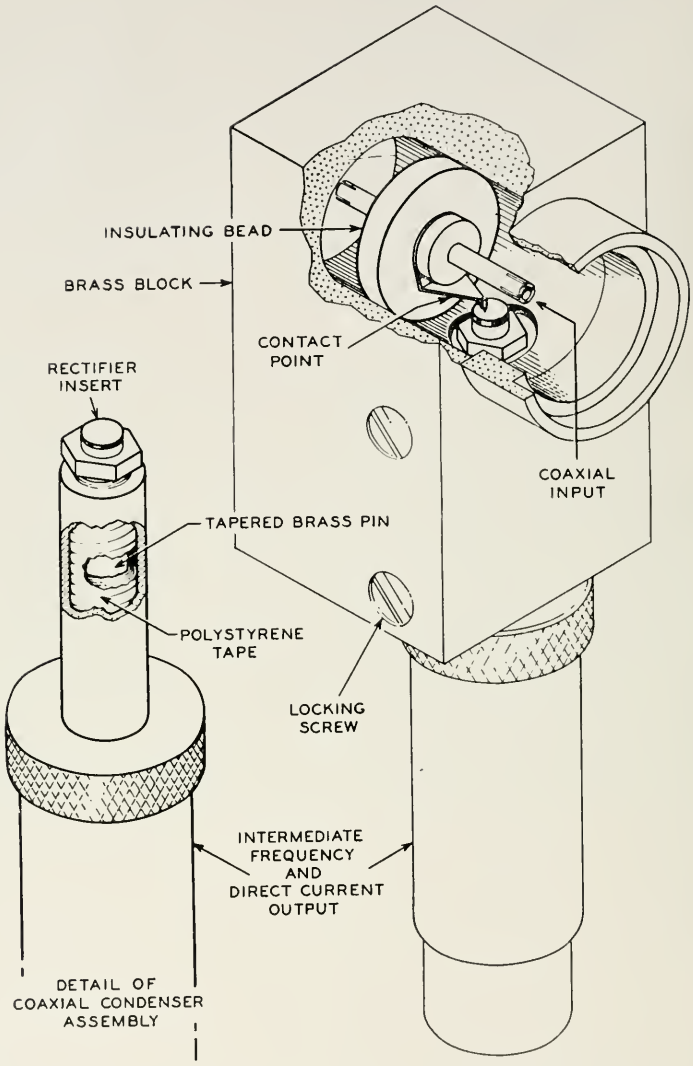


Fig. 3—Schematic diagram of one of the early crystal converter blocks.

The inserts and points in appropriate mounting blocks were widely used in centimeter wave investigations prior to 1940.[2] The principal laboratory uses were in frequency converter circuits in receivers, and as radio fre-

[2] G. C. Southworth and A. P. King, "Metal Horns as Directive Receivers of Ultra-Short Waves," *Proc. I. R. E.* v. 27, pp. 95–102, 1939; Carl R. Englund, "Dielectric Constants and Power Factors at Centimeter Wave Lengths," *Bell Sys. Tech. Jour.*, v. 23, pp. 114–129, 1944; Brainerd, Koehler, Reich, and Woodruff, "Ultra High Frequency Techniques," D. Van Nostrand Co., Inc., 250–4th Avenue, New York, 1942.

quency instrument rectifiers. They were also used to a relatively minor extent in some of the early radar test equipment. Moreover, the availability of these devices and the knowledge of their properties as microwave converters tended to focus attention on the potentialities of radar designs employing crystal rectifiers in the receiver's frequency converter. Similarly, the techniques established for preparation of the inserts tended to orient subsequent manufacturing process developments. For example, the methods now generally used for preparing silicon ingots, for cutting the rectifying element from the ingot with diamond saws, and for forming the

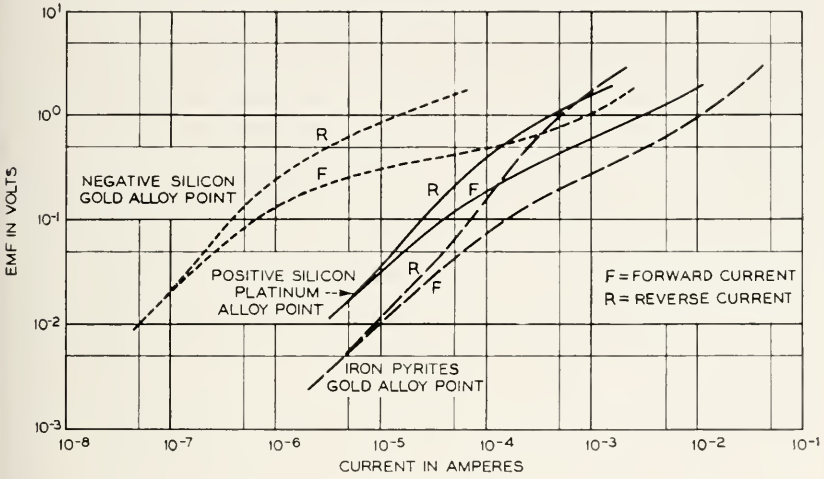


Fig. 4—Direct-current characteristics of silicon and iron pyrite rectifiers fabricated as inserts, 1939.

back contact to the rectifying element by electroplating procedures, are still essentially similar to the techniques used for preparing the inserts in 1939. As a contribution to the defense research effort, this basic information, with various samples and experimental assemblies, was made available to governmental agencies for dissemination to authorized domestic and foreign research establishments.

Development of the Ceramic Type Cartridge Structure

The block rectifier structure previously described was well adapted to various laboratory needs because of its flexibility, but for large scale utilization certain limitations are evident. Not only was it necessary that the parts be accurately machined, but also the adjustment of the rectifier in

the block structure required considerable skill. With recognition of the military importance of silicon crystal rectifiers, effort was intensified in the development of standardized structures suitable for commercial production.

In the 1940–1941 period, contributions to the design of silicon crystal rectifiers were made by British workers as a part of their development of new military implements. For these projected military uses, the problem of replacement and interchangeability assumed added importance. The design trend was, therefore, towards the development of a cartridge type structure with the electrical adjustment fixed during manufacture, so that the unit could be replaced easily in the same manner as vacuum tubes.

In the latter part of 1941 preliminary information was received in this country through National Defense Research Committee channels on a rectifier design originating in the laboratories of the British Thomson-Houston Co., Ltd. A parallel development of a similar device was begun in various American laboratories, including the Radiation Laboratory at the Massachusetts Institute of Technology, and Bell Telephone Laboratories. In the work at Bell Laboratories, emphasis was placed both on development of a structure similar to the British design and on exploration and test of various new structures which retained the features of socket interchangeability but which were improved mechanically and electrically.

In the work on the ceramic cartridge, the external features of the British design were retained for reasons of mechanical standardization but a number of changes in process and design were made both to improve performance and to simplify manufacture. To mention a few, the position of the silicon wafer and the contact point were interchanged because measurements indicated that an improvement in performance could thereby be obtained. To obviate the necessity for searching for active spots on the surface of the silicon and to improve performance, fused high purity silicon was substituted for the "commercial" silicon then employed by the British. The rectifying element was cut from the ingots by diamond saws, and carefully polished and etched to develop optimum rectification characteristics. Similar improvements were made in the preparation of the point or "cats whisker", replacing hand operations by machine techniques. To protect the unit from mechanical shock and the ingress of moisture, a special impregnating compound was developed which was completely satisfactory even under conditions of rapid changes in temperature from $-40°$ to $+70°$C. All such improvements were directed towards improving quality and establishing techniques for mass production.

In this early work time was at a premium because of the need for prompt standardization of the design in order that radar system designs might in

turn be standardized, and that manufacturing facilities might be established to supply adequate quantities of the device. The development and initial production of the device was accomplished in a short period of time. This was possible because process experience had been acquired in the insert development, and centimeter wave measurements techniques and facilities were then available to measure the characteristics of experimental units at the operating frequency. By December 1941, a pattern of manufacturing techniques had been established so that production by the Western Electric Company began shortly thereafter. This is believed to have been the first commercial production of the device in this country.

As a result of the basic information on centimeter wave measurements techniques which was available from earlier microwave research at the Holmdel Radio Laboratory, it was possible also, at this early date, to propose to the Armed Services that each unit be required to pass an acceptance test consisting of measurement of the operating characteristics at the intended operating frequency. This plan was adopted and standard test methods devised for production testing. Considering the complexity of centimeter wave measurements, this was an accomplishment of some magnitude and was of first importance to the Armed Services because it assured by direct measurement that each unit would be satisfactory for field use.

The cartridge structure resulting from these developments and meeting the international dimensional standards is shown in Fig. 5. It consists of two metal terminals separated by an internally threaded ceramic insulator. The rectifying element itself consists of a small piece of silicon (p-type) soldered to the lower metal terminal or base. The contact spring or "cats whisker" is soldered into a cylindrical brass pin which slides freely into an axial hole in the upper terminal and may be locked in any desired position by set screws. The spring itself is made from tungsten wire of an appropriate size, formed into an S shape. The free end of the wire, which in a finished unit engages the surface of the silicon and establishes rectification, is formed to a cone-shaped configuration in order that the area of contact may be held at the desired low value.

The silicon elements used in the rectifiers are prepared from ingots of fused high purity silicon. Alloying additions are made to the melt when required to adjust the electrical resistivity of the silicon to the value desired. The ingots are then cut and the silicon surfaces prepared and cut into small pieces approximately 0.05 inch square and 0.02 inch thick suitable for use in the rectifiers. The contact springs are made from tungsten wire, gold plated to facilitate soldering. Depending upon the application, the wires

may be 0.005 inch, 0.0085 inch, or 0.010 inch in diameter. After forming the spring to the desired shape, the tip is formed electrolytically.

In assembling the rectifier cartridge, the two end terminals, consisting of the base with the silicon element soldered to it, and the top detail containing the contact spring, are threaded into the ceramic tube so that the free end of the spring does not engage the silicon surface. An adhesive

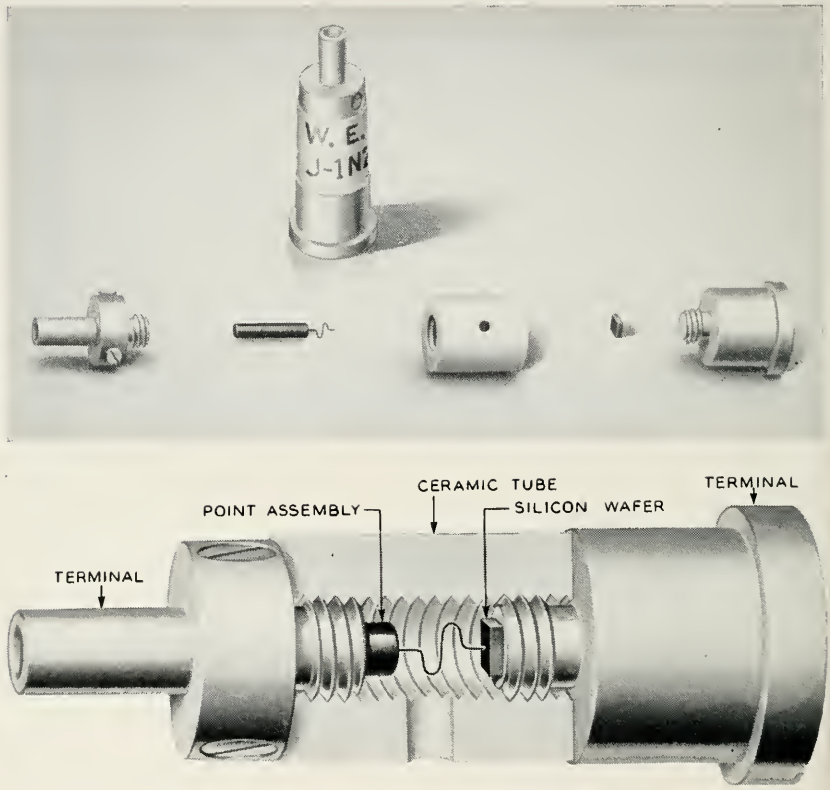


Fig. 5—Ceramic cartridge rectifier structure and parts.
Overall length of assembled rectifier is approximately $\frac{3}{4}$ inch.

is employed to secure the parts firmly to the ceramic. The rectifier is then "adjusted" by bringing the point into engagement with the silicon surface and establishing optimum electrical characteristics. Finally the unit is impregnated with a special compound to protect it from moisture and from damage by mechanical shock. Units so prepared are then ready for the final electrical tests.

The adjustment of the rectifier is an interesting operation for at this

stage in the process the rectification action is developed, and to a considerable degree, controlled. If the point is brought into contact with the silicon surface and a small compressional deflection applied to the spring, direct-current measurements will show a moderate rectification represented by the passage of more current at a given voltage in the forward direction than in the reverse. If the side of the unit is now tapped sharply by means of a small hammer, the forward current will be increased, and, at the same time, the reverse current decreased.[3] With successive blows the reverse current is reduced rapidly to a constant low value while the forward current increases, but at a diminishing rate, until it also becomes relatively constant. The magnitude of the changes produced by this simple operation is rather surprising. The reverse current at one volt seldom decreases by less than a factor of 10 and frequently decreases by as much as a factor of 100, while the forward current at one volt increases by a factor of 10. Paralleling these changes are improvements in the high-frequency properties, the conversion loss and noise both being reduced. The tapping operation is not a haphazard searching for better rectifying spots, for with a given silicon material and mechanical assembly the reaction of each unit to tapping is regular, systematic and reproducible. The condition of the silicon surface also has a pronounced bearing on "tappability" for by modifications of the surface it is possible to produce, at will, materials sensitive or insensitive in their reaction to the tapping blows.

In the development of the compounds for filling the rectifier, special problems were met. For example, storage of the units for long periods of time under either arctic or tropical conditions was to be expected. Also, for use in air-borne radars operating at high altitudes, where equipment might be operated after a long idle period, it was necessary that the units be capable of withstanding rapid heating from very low temperatures. The temperature range specified was from $-40°$ to $+70°$C. Most organic materials normally solid at room temperature, as the hydrocarbon waxes, are completely unsuitable, as the excessive contraction which occurs at low temperatures is sufficient to shift the contact point and upset the precise adjustment of the spring. Nor are liquids satisfactory because of their tendency to seep from the unit. However, special gel fillers, consisting of a wax dispersed in a hydrocarbon oil, were devised in Bell Telephone Laboratories to meet the requirements, and were successfully applied by the leading manufacturers of crystal rectifiers in this country. Materials of a similar nature, though somewhat different in composition, were also used subsequently in Britain. Further improvements in these compounds have been made recently, extending the temperature range 10°C at low

[3] Southworth and King; loc. cit.

temperatures and about 30°C at high temperatures in response to the design trend towards operation of the units at higher temperatures. The units employing this compound may, if desired, be repeatedly heated and cooled rapidly between −50°C and +100°C without damage.

Use of the impregnating compound not only improves mechanical stability but prevents ingress or absorption of moisture. Increase of humidity would subject the unit not only to changes in electrical properties such as variation in the radio frequency impedance, but also to serious corrosion, for the galvanic couple at the junction would support rapid corrosion of the metal point. In fact, with condensed moisture present in unfilled units corrosion can be observed in 48 hours. For this reason alone, the development of a satisfactory filling compound was an important step in the successful utilization of the units by the Armed Services under diverse and drastic field conditions.

TABLE I

Shelf Aging Data on Silicon Crystal Rectifiers of the Ceramic Cartridge Design

Storage Conditions	Initial Values		Values After Storage for 7 Months	
	Conversion Loss (Median) (L)	Noise Ratio (Median) (N_r)	Conversion Loss (median) (L)	Noise Ratio (median) (N_r)
	db	*db*	*db*	*db*
75°F. 65% Relative Humidity	6.8	3.9	6.7	4.3
110°F. 95% Relative Humidity	6.9	3.9	6.9	4.3
−40°C	7.0	3.9	6.8	3.9

The large improvement in stability achieved in the present device as compared with the older crystal detectors may be attributed to the design of the contact spring, correct alignment of parts in manufacture and to the practice of filling the cavity in the unit with the gel developed for this purpose. Considering the apparently delicate construction of the device, the stability to mechanical or thermal shock achieved by these means is little short of spectacular. Standard tests consist of dropping the unit three feet to a wood surface, immersing in water, and of rapidly heating from −40 to 70°C. None of these tests impairs the quality of the unit. Similarly the unit will withstand storage for long periods of time under adverse conditions. Table I summarizes the results of tests on units which were stored for approximately one year under arctic (−40°), tropical (114°F—95% relative humidity), and temperate conditions. Though minor changes in the electrical characteristics were noted in the accelerated tropical test, none of the units was inoperative after this drastic treatment.

Development of the Shielded Rectifier Structure

Rectifiers of the ceramic cartridge design, though manufactured in very large quantities and widely and successfully used in military apparatus, have certain well recognized limitations. For example, they may be accidentally damaged by discharge of static electricity through the small point contact in the course of routine handling. If one terminal of the unit is held in the hand and the other terminal grounded, any charge which may have accumulated will be discharged through the small contact. Since such static charges result in potential differences of several thousand volts it is understandable that the unit might suffer damage from the discharge. Although damage from this cause may be avoided by following a few simple precautions in handling, the fact that such precautions are needed constitutes a disadvantage of the design.

Certain manufacturing difficulties are also associated with the use of the threaded insulator. The problem of thread fit requires constant attention. Lack of squareness at the end of the ceramic cylinder or lack of concentricity in the threaded hole tends to cause an undesirable eccentricity or angularity in the assembled unit which can be minimized only by rigid inspection of parts and of final assemblies. At the higher frequencies (10,000 megacycles), uniformity in electrical properties, notably the radio frequency impedance, requires exceedingly close control of the internal mechanical dimensions. In the cartridge structure where the terminal connections are separated by a ceramic insulating member, the additive variations of the component parts make close dimensional control inherently difficult.

To eliminate these difficulties the shielded structure, shown in Fig. 6, was developed. In this design the rectifier terminates a small coaxial line. The central conductor of the line, forming one terminal of the rectifier, is molded into an insulating cylinder of silica-filled bakelite, and has spot welded to it a 0.002-inch diameter tungsten wire spring of an offset C design. The free end of the spring is cone shaped. The rectifying element is soldered to a small brass disk. Both the disk, holding the rectifying element, and the bakelite cylinder, holding the point, are force-fits in the sleeve which forms the outer conductor of the rectifier. By locating the bakelite cylinder within the sleeve so that the free end of the central conductor is recessed in the sleeve, the unit is effectively protected from accidental static damage as long as the holder or socket into which the unit fits is so designed that the sleeve establishes electrical contact with the equipment at ground potential before the central conductor. The sleeve also shields the rectifying contact from effects of stray radiation.

The radio frequency impedance of the shielded unit can be varied within certain limits by modifying the diameter of the central conductor. For

example, in the 1N26 unit, which was designed for use at frequencies in the region of 24,000 megacycles, a small metal slug fitting over the central conductor makes it possible to match a coaxial line having a 65-ohm surge impedance. For certain circumstances this modification in design is advantageous, while in others it is a disadvantage because the matching slug is effective only over a narrow range of frequencies.

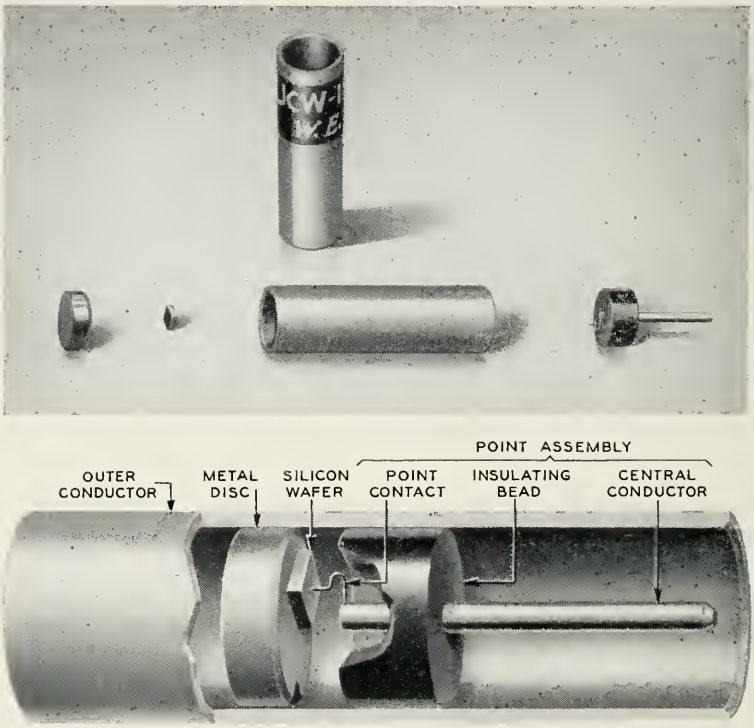


Fig. 6—Shielded rectifier structure and parts. Overall length of assembled rectifier is approximately $\frac{3}{4}$ inch.

The shielded structure was developed in 1942 and since it was of a simplified design with reduced hazard of static damage, it was proposed to the Armed Services for standardization in June of that year. However, because of the urgency of freezing the design of various radars and because the British had already standardized on the outline dimensions of the ceramic type cartridge, Fig. 5, the Services did not consider it advantageous to standardize the new structure when first proposed. In deference to this international standardization program, plans for the manufacture of this

structure were held in abeyance during 1942 and 1943. However, an opportunity for realizing the advantages inherent in the shielded design was afforded later in the war and a sufficient quantity of the units was produced to demonstrate its soundness. As anticipated from the constructional features, marked uniformity of electrical properties was obtained.

Types, Applications, and Operating Characteristics

Various rectifier codes, engineered for specific military uses, were manufactured by Western Electric Company during the war. These are listed in Table II. The units are designated by RMA type numbers, as 1N21, 1N23, etc., depending upon their properties and the intended use. Letter suffixes, as 1N23A, 1N23B, indicate successively more stringent performance requirements as reflected in lower allowable maxima in loss and noise ratio, and, usually, more stringent power proof-tests. In general, different codes are provided for operation in the various operating frequency ranges. For example, the 1N23 series is tested at 10,000 megacycles while the 1N21 series is tested at 3,000 megacycles and the 1N25 at 1000 megacycles, approximately. Since higher transmitter powers are frequently employed at the lower frequencies, somewhat greater power handling ability is provided in units for operation in this range.

One of the more important uses of silicon crystal rectifiers in military equipment was in the frequency converter or first detector in superheterodyne radar receivers. This utilization was universal in microwave receivers. In this application the crystal rectifier serves as the non-linear circuit element required to generate the difference (intermediate) frequency between the radio frequency signal and the local oscillator. The intermediate frequency thus obtained is then amplified and detected in conventional circuits. As the crystal rectifier is normally used at that point in the receiving circuit where the signal level is at its lowest value, its performance in the converter has a direct bearing on the overall system performance. It was for this reason that continued improvements in the performance of crystal rectifiers were of such importance to the war effort.

For the converter application, the signal-to-noise properties of the unit at the operating frequency, the power handling ability, and the uniformity of impedance are important factors. The signal-to-noise properties are measured as conversion loss and noise ratio. The loss, L, is the ratio of the available radio frequency signal input power to the available intermediate frequency output power, usually expressed in decibels. The noise ratio, N_R, is the ratio of crystal output noise power to thermal (KTB) noise power. The loss and noise ratio are fundamental properties of the

TABLE II

Performance Specifications—Silicon Crystal Rectifiers

Code	W.E. Spec. No.	Structure	Nominal Operating Frequency	Test Power	Max. Conversion Loss	Max. Noise Ratio[7]	Int. Freq. Impedance	Min. Figure of Merit	Low-Level Resistance	Proof Test Method, Note	Test Level	Normal Use
			mega-cycles	*milli-watts*	*db*		*ohms*		*ohms*			
1N21	D-163366	Cartridge	3000	0.4	8.5	4.0						Converter
1N21A	D-168355	"	"	0.4	7.5	3.0				5	0.3 erg	"
1N21B	D-169113	"	"	0.5	6.5	2.0	375[1]			5	2.0 ergs	"
1N23	D-168356	"	10000	1.0	10.0	3.0	"			5	0.3 erg	"
1N23A	D-169304	"	"	1.0	8.0	2.7	"			5	1.0 erg	"
1N23B	D-169305	"	"	1.0	6.5	2.7	"			5	0.3 erg	"
1N25	D-170247	"	1000	1.25	8.0	2.5	100–400[2]			6	$12.5v^3$ $6.\omega^4$	"
1N26	D-168707	Shielded	24000	1.0	8.5	2.5	300–600[2]			5	0.1 erg	"
1N27	D-172981	Cartridge	3000	.005 max.				60	4000 max. at 1 mv			Low-Level Detector
1N28	D-168030	"	"	0.4	7.0	2.0	300[1]			5	5.0 erg	Converter
1N31	D-171936	Shielded	10000	.005 max.				55	6000–23,000 at 5 mv max. 20–25°C.	6	$1.5v^3$ $25.\omega^4$	Low-Level Detector

[1] Average values under conditions of production test—not limited directly by specifications.
[2] Values limited by specifications.
[3] Pulse generator voltage.
[4] Pulse generator impedance.
[5] Single D.C. spike.
[6] Multiple square pulse.
[7] Numerical ratio.

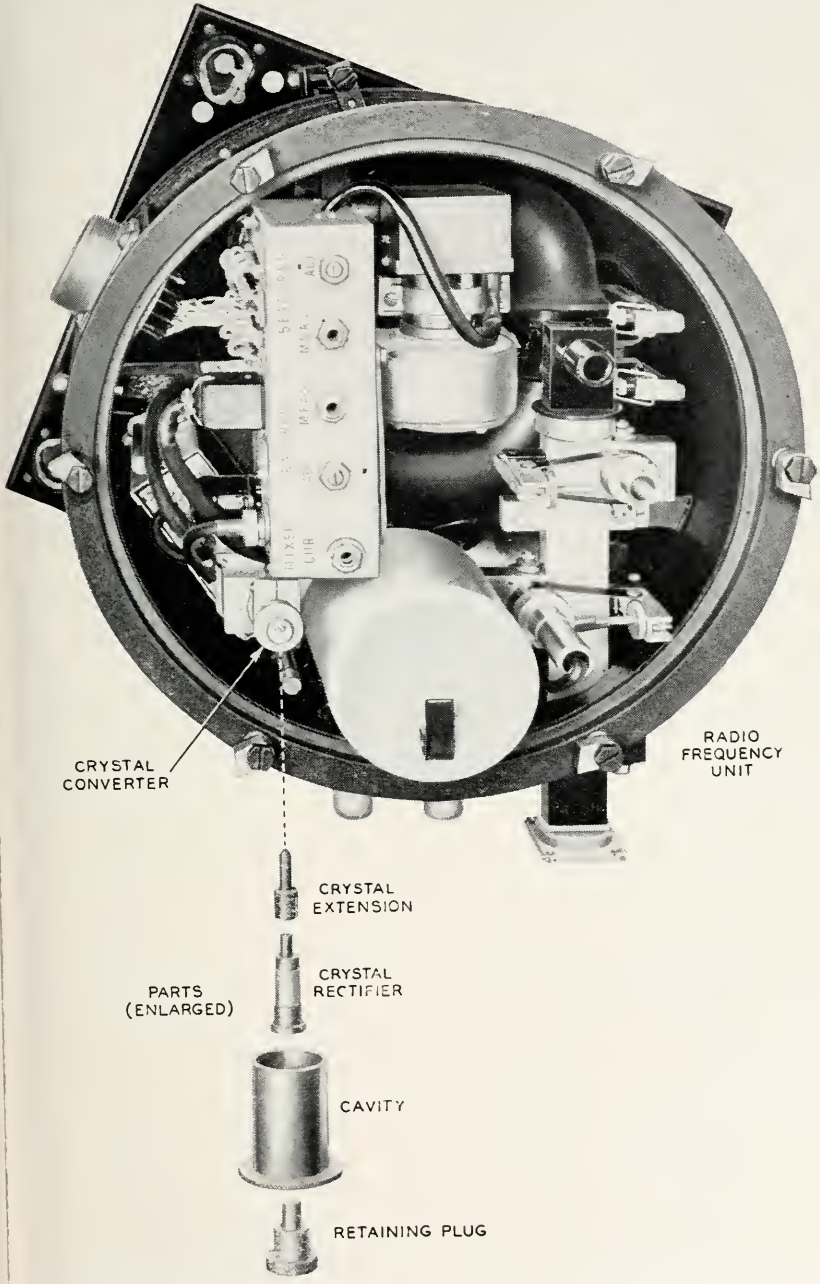


Fig. 7—Converter for wave guide circuits as installed in the radio frequency unit of the AN/APQ13 radar system. This was standard equipment in B-29 bombers for radar bombing and navigation.

converter. From these data and other circuit constants, the designer may calculate[4] expected receiver performance.

For operation as converters,[5] crystal rectifiers are employed in suitable holders. These may be arranged for use with either coaxial line or wave guide circuits, depending upon the application. Figure 7 shows a converter for wave guide circuits installed in the radio frequency unit of an air-borne radar system. A typical converter designed for use with coaxial lines is shown in the photograph Fig. 8. A schematic circuit of this converter is shown in Fig. 9. In such circuits the best signal-to-noise ratio is realized when an optimum amount of beating oscillator power is supplied. The optimum power depends, in part, on the properties of the rectifier itself, and, in part, on other circuit factors as the noise figure of the intermediate

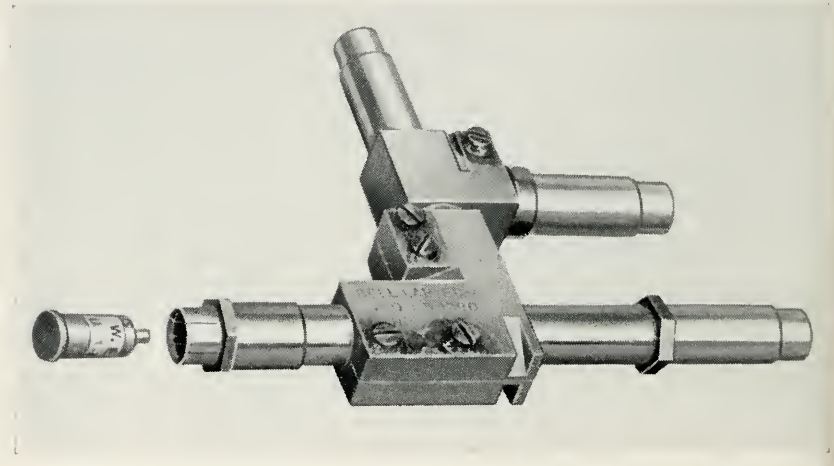


Fig. 8—Converter for use at 3000 megacycles. The crystal rectifier is located adjacent to its socket in the converter.

frequency amplifier. For a well designed intermediate frequency amplifier with a noise figure of about 5 decibels, the optimum beating oscillator power is such that between 0.5 and 2.0 milliamperes of rectified current flows through the rectifier unit. Under these conditions and with the unit matched to the radio frequency line, the beating oscillator power absorbed by the unit is about one milliwatt. For intermediate frequency amplifiers

[4] The quantities L and N_R are related to receiver performance by the relationship

$$F_R = L(N_R - 1 + F_{IF})$$

where F_R is the receiver noise figure and F_{IF} is the noise figure of the intermediate frequency amplifier. All terms are expressed as power ratios. A rigorous definition of receiver noise figure has been given by H. T. Friis "Noise Figures of Radio Receivers," *Proc. I. R. E.*, vol. 32, pp. 419–422; July, 1944.

[5] C. F. Edwards, "Microwave Converters," presented orally at the Winter Technical Meeting of the *I. R. E.*, January 1946 and submitted to the *I. R. E.* for publication.

with poorer noise figures, the drive for optimum performance is higher than the figures cited above. Conversely, for intermediate frequency amplifiers with exceptionally low noise figures, optimum performance is obtained with lower values of beating oscillator drive. If desired, somewhat higher currents than 2.0 milliamperes may be employed without damage to the crystal.

The impedance at the terminals of a converter using crystal rectifiers, both at radio and intermediate frequencies, is a function not only of the rectifier unit, but also of the circuit in which the unit is used and of the

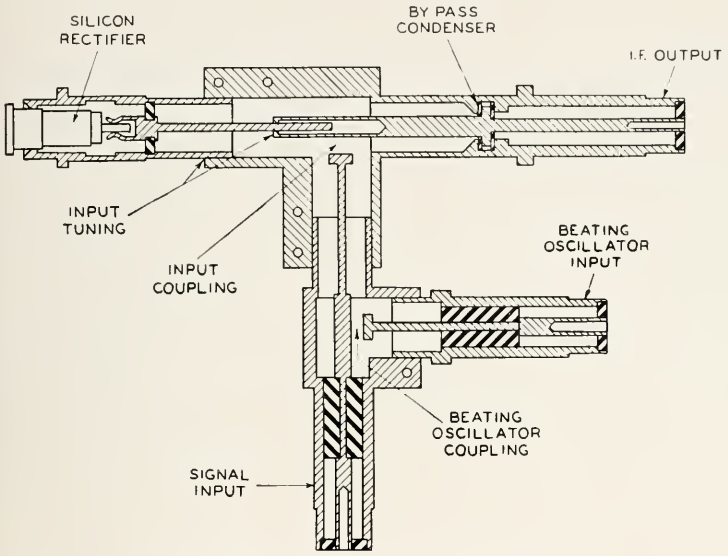


Fig. 9—Schematic diagram of crystal converter.

power level at which it is operated. Consequently the specification of an impedance for a crystal rectifier is of significance only in terms of the circuit in which it is measured. Since the converters used in the production testing of crystal rectifiers are not necessarily the same as those used in the field, and since in addition there are frequently several converter designs for the same type of unit, a specification of crystal rectifier impedance in production testing can do little more than select units which have the same impedance characteristic in the production test converter. The impedances at the terminals of two converters of different design but using the same crystal rectifier may vary by a factor of 3 or even more, with the intermediate frequency impedance generally varying more drastically than the radio frequency impedance. The variation is also a function of the con-

version loss. Crystals with large conversion losses are less susceptible to impedance changes from reactions in the radio frequency circuit than are low conversion loss units.

The level of power to which the rectifiers can be subjected depends upon the way in which the power is applied. The application of an excessive amount of power or energy results in the electrical destruction of the unit by rupture of the rectifying material. Experimental evidence indicates that the electrical failure may be in one of three categories. The total energy of an applied pulse is responsible for the impairment when the pulse length is shorter than 10^{-7} seconds, the approximate thermal time constant of the crystal rectifier as given by both measurement and calculation. For pulse lengths of the order of 10^{-6} seconds the peak power in the pulse is the determining factor, and for continuous wave operation the limitation is in the average power.

In performance tests in manufacture all units for which burnout tolerances are specified are subjected to proof-tests at levels generally comparable with those which the unit may occasionally be expected to withstand in actual use, but greater than those to be employed as a design maximum. The power or energy is applied to the unit in one of two types of proof-test equipment. The multiple, long time constant (of the order of 10^{-6} seconds) pulse test is applied to simulate the plateau part of a radar pulse reaching the crystal through the gas discharge transmit-receive switch.[6] This test uses an artificial line of appropriate impedance triggered at a selected repetition rate for a determined length of time. The power available to the unit is computed from the usual formula,

$$P = \frac{V^2}{4Z},$$

where P is the power in watts, V is the potential in volts to which the pulse generator is charged, and Z is the impedance in ohms of the pulse generator. In general, where this test is employed, a line is used which matches the impedance of the unit under test at the specified voltage.

The second type of test is the single discharge of a coaxial line through the unit to simulate a radar pulse spike reaching the crystal before the transmit-receive switch fires. The pulse length is of the order of 10^{-9} second. The energy in the test spike may be computed from the relation

$$E = \frac{10^7}{2} CV^2,$$

where E is the energy in ergs, C the capacity of the coaxial line in farads, and V the potential in volts to which the line is charged.

[6] A. L. Samuel, J. W. Clark, and W. W. Mumford, "The Gas Discharge Transmit-Receive Switch," *Bell Sys. Tech. Jour.*, v. 25 No. 1, pp. 48-101, Jan. 1946.

Specification proof-test levels are, of course, not design criteria. Since the units are generally used in combination with protective devices, such as the transmit-receive switch, it is necessary to conduct tests in the circuits of interest to establish satisfactory operating levels.

In general, however, the units may be expected to carry, without deterioration, energy of the order of a third of that used in the single d-c spike proof-test or peak powers of a magnitude comparable with that used in the multiple flat-top d-c pulse proof-test. The upper limit for applied continuous wave signals has not been determined accurately, but, in general, rectified currents below 10 milliamperes are not harmful when the self bias is less than a few tenths of a volt.

The service life of a crystal rectifier will depend completely upon the conditions under which it is operated and should be quite long when its ratings are not exceeded. During the war, careful engineering tests conducted on units operating as first detectors in certain radar systems revealed no impairment in the signal-to-noise performance after operation for several hundred hours. A small group of 1N21B units showed only minor impairments when operated in laboratory tests for 100 hours with pulse powers (3000 megacycles) up to 4 watts peak available to the unit under test.

Another important military application of silicon crystal rectifiers was as low-power radio frequency rectifiers for use in wave meters or other items of radar test equipment. Here the rectification properties of the unit at the operating frequency are of primary interest. Since units which are satisfactory as converters also function satisfactorily as high-frequency rectifiers special types were not required for this application.

Units were also used in military equipment as detectors to derive directly the envelope of a radio frequency signal received at low power levels. These signals were modulated usually in the video range. The low-level performance is a function of the resistance at low voltages and the direct-current output for a given low-power radio frequency input. These may be combined to derive a figure of merit which is a measure of receiver performance.[7]

Typical direct-current characteristics of the silicon rectifiers at temperatures of $-40°$, 25° and 70°C are given in Fig. 10. It will be noted in these curves that both the forward and reverse currents are decreased by reducing the temperature and increased by raising the temperature. The reverse current changes more rapidly with temperature than the forward current, however, so that the rectification ratio is improved by reducing the temperature, and impaired by raising the temperature. The data shown are for typical units of the converter type. It should be emphasized, however,

[7] R. Beringer, Radiation Laboratory Report No. 61–15, March 16, 1943.

that by changes in processing routines the direct-current characteristics shown in Fig. 10 may be modified in a predictable manner, particularly with respect to absolute values of forward current at a particular voltage.

Modern Rectifier Processes

When the development of the type 1N21 unit was undertaken, the scientific and engineering information at hand was insufficient to permit intentional alteration or improvement in electrical properties of the rectifier. In these early units, the control of the radio frequency impedance, power handling ability and signal-to-noise ratio left much to be desired. Within a short time, some improvements in performance were realized by process improvements such as the elimination of burrs and irregularities from the point contact to reduce noise. Substantial improvements were not obtained,

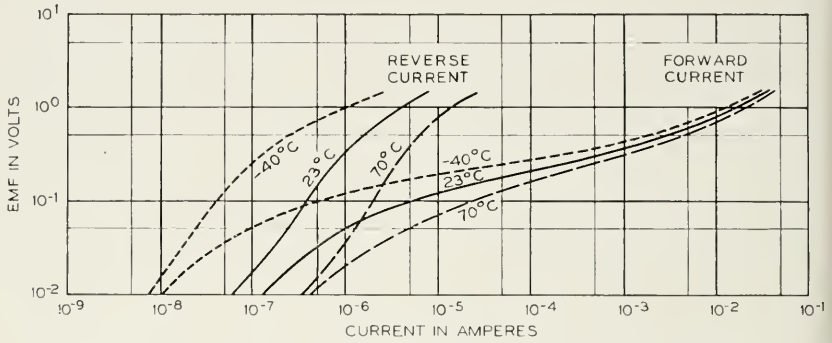


Fig. 10—Direct-current characteristics of P-type silicon crystal rectifier at various temperatures.

however, until certain improved materials, processes, and techniques were developed.

In the engineering development of improved crystal rectifier materials and processes, basic data have been acquired which make it possible to alter the properties of the rectifier in a predictable manner so that the units may now be engineered to the specific electrical requirements desired by the circuit designer in much the same manner as are modern electron tubes. This has led not only to improvements in performance but also to a diversification in types and applications.

The simplified equivalent circuit for the point contact rectifier, shown in Fig. 11, provides a basis for consideration of the various process features. In Fig. 11, C_B represents the electrical capacitance at the boundary between the point contact and the semi-conductor, R_B the non-linear resistance at this boundary, and R_S is the spreading resistance of the semi-conductor

proper, that is the total ohmic resistance of the silicon to current through the point. The capacitance C_B being shunted across the rectifying boundary, decreases the efficiency of the device by its by-pass action because the current through it would be dissipated as heat in the resistance R_S. Losses from this source increase rapidly with increased frequency because of the enhanced by-pass action. It would appear, therefore, that to improve efficiency it would be important to minimize both R_S and C_B by some method such as reducing the area of the rectifying contact and lowering the body resistance of the silicon employed. For a given silicon material, the impedances desired for reasons of circuitry and considerations of mechanical stability place a limit on the extent to which performance may be improved by reducing the contact area. R_S may be reduced by using silicon of lower resistivity, but this generally results in poorer rectification. This impairment is due apparently to some subtle change in the properties of the rectifying junction resulting from decreasing the specific resistance of the silicon material.

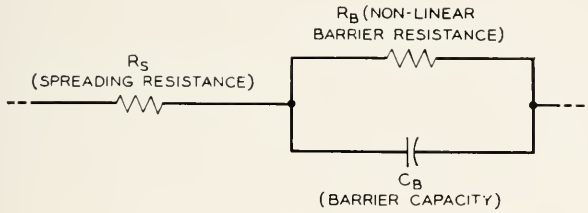


Fig. 11—Simplified equivalent circuit of crystal rectifier.

The answer to this apparent dilemma lies in the application of an oxidizing heat treatment to the surface of the semi-conductor. This process derives from researches conducted independently in this country and in Britain, though there was considerable interchange of information between various interested laboratories. In the oxidizing treatment, apparently the impurities in the silicon which contribute to its conductivity diffuse into the adhering silica film, thereby depleting impurities from the surface of the silicon. When the oxide layer is then removed by solution in dilute hydrofluoric acid, the underlying silicon layer is exposed and remains intact as the acid does not readily attack the silicon itself.

Since decreasing the impurity content of a semi-conductor increases its resistivity, the silicon surface has higher resistivity after the oxidizing treatment than before. Thus by oxidation of the surface of low resistance silicon it is possible to secure the enhanced rectification associated with the high resistance surface layer, while by virtue of the lower resistivity of the underlying material the I^2R losses through R_S are reduced.

In actual practice the properties of the rectifier are governed by the resistivity of the silicon material, the contact area, and the degree of oxidation of the surface. By the controlled alteration of these factors units may be engineered for specific applications. The body resistance of the silicon is controlled by the kind and quantity of the impurities present. Aluminum, beryllium or boron may be added to purified silicon to reduce its resistivity to the desired level. Boron is especially effective for this purpose, the quantity added usually being less than 0.01 per cent. As little as 0.001 per cent has a very pronounced effect upon the electrical properties. The contact area is determined by the design of contact spring employed and the deflection applied to it in the adjustment of the rectifier. The degree of oxidation is controlled by the time and temperature of the treatment and the atmosphere employed.

In the development of the present rectifier processes, certain experimental relationships were obtained between the performance and the contact area on the one hand, and the power handling ability and contact area on the other. These show the manner in which the processes should be changed to produce a desired change in properties. For example, Fig. 12 shows the relationship between the spring deflection applied to a unit and the conversion loss at a given frequency. The apparent contact area, (i.e., the area of the flattened tip of the spring in contact with the silicon surface, as measured microscopically) also increases with increasing spring deflection. It will be seen in Fig. 12 that for a given silicon material, the conversion loss at 10,000 megacycles increases rapidly with the contact area. The curves tend to reach constant loss values at the higher spring deflections. It is believed that this may be ascribed to the fact that for a given spring size and form, the increment in contact area obtained by successive increments in spring deflection would diminish and finally become zero after the elastic limit of the spring is exceeded.

The losses plotted in Fig. 12 were measured on a tuned basis, that is, the converter was adjusted for maximum intermediate frequency output at a fixed beating oscillator drive for each measurement. Were these measurements made on a fixed tuned basis, that is, with the converter initially adjusted for maximum intermediate frequency output for a unit to which the minimum spring deflection is applied, and the units with larger deflections then measured without modification of the converter adjustment, even greater degradation in conversion loss than that shown in Fig. 12 would be observed. This results from the dependence of the radio frequency impedance upon the contact area. In loss measurements made on the tuned basis, changes in the radio frequency impedance occasioned by the changes in the contact area do not affect the values of mismatch loss obtained, while on the

fixed tuned basis they would result in an increase in the apparent loss because of the mismatch of the radio frequency circuits.

While the conversion loss is degraded by increasing the contact area, the power handling ability[8] of the rectifiers is improved, as shown in Fig. 13.

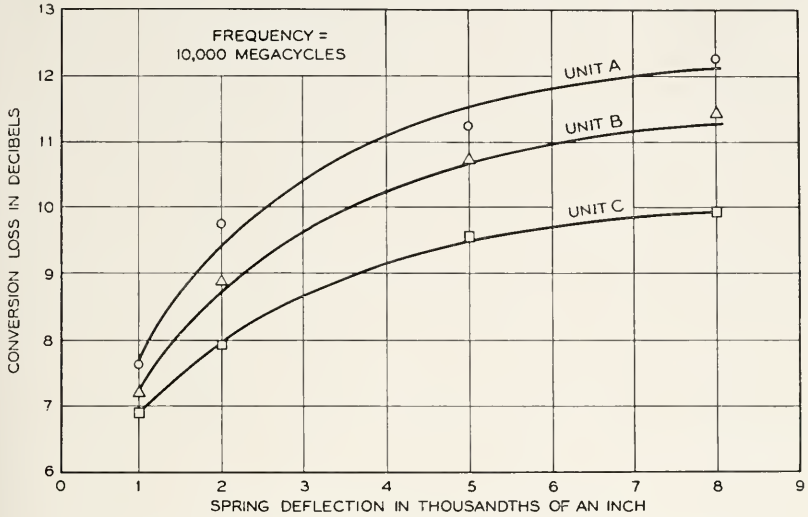


Fig. 12—Relationship between spring deflection and conversion loss in silicon crystal rectifiers.

This is not surprising because the larger area contact gives a wider current distribution and thus minimizes the localized heating effects near the contact. Generally, therefore, in the development of units for operation at a

[8] The measurement of power handling ability of crystal rectifiers by application of radio frequency power is complicated by the fact that the impedance of the unit under test varies with power level. If a unit is matched in a converter at a low-power level and power at a higher level is then applied, not all of the power available is absorbed by the unit but a portion of it is reflected (due to the change in impedance). This factor has been called the self protection of the unit and it necessitates the distinction between the power absorbed by and the power available to the unit under test. The data for Fig. 13 were acquired by first matching the unit in converters at low powers (about 0.3 milliwatts CW 3000 mcs) and then exposing it for a short period to successively higher levels of pulse power of square wave form of 0.5 microseconds width at a repetition rate of 2000 pulses per second, measuring the loss and noise ratio after each power application. The power handling ability is then expressed as the available peak power required to cause a 3 db impairment in the conversion loss or the receiver noise figure. This method was employed because in radar receivers the units are matched for low-power levels. In this respect the method simulates field operating conditions, but the "spike" of radar pulses is absent.

The increase in power handling ability with increasing area shown in Fig. 13 is confirmed by similar measurements with radio frequency pulse power with the unit matched at high-level powers, by direct-current tests, and by simple 60-cycle continuous wave tests. The magnitude of the increase depends, however, upon the particular method employed for measurement.

given frequency, a compromise must be effected between these two important performance factors. Because of increased condenser by-pass action a smaller area must be used to obtain a given conversion loss at a higher frequency. For this reason the power handling ability of units designed for use at the higher frequencies is somewhat less than that of the lower-fre-

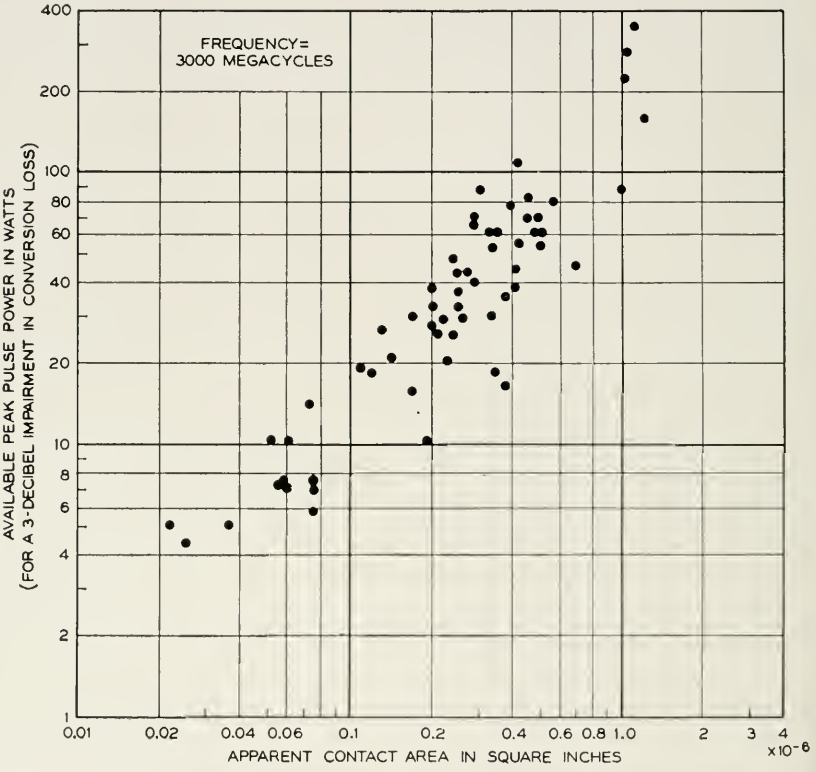


Fig. 13—Correlation between power handling ability measured with microsecond radio frequency pulses and contact area in silicon crystal rectifiers.

quency units because emphasis has been placed upon achieving a given signal-to-noise performance in each frequency band.

Use of the improved materials and processes produced rather large improvements in the d-c rectification ratio, conversion loss, noise, power handling ability, and uniformity. Typical direct-current rectification characteristics of units produced by both the old and the new processes are shown in Fig. 14. These curves show that reverse currents at one volt were decreased by a factor of about 20 while the forward currents were increased by

a factor of approximately 2.5 giving a net improvement in rectification ratio of 50 to 1. The parallel improvement in receiver performance resulting from process improvements is shown in Fig. 15. A comparison in power handling

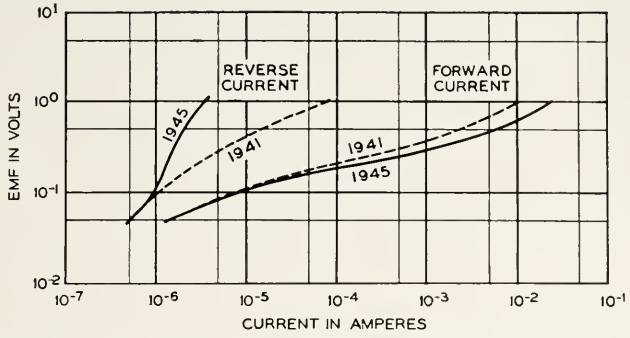


Fig. 14—Improvement in the direct-current rectification characteristics of silicon crystal rectifiers in a four-year period.

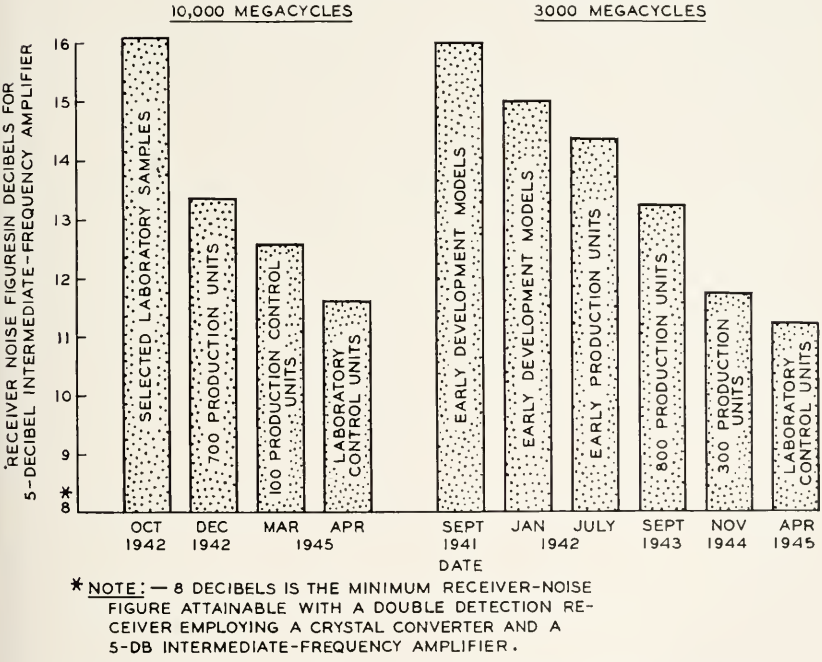


Fig. 15—Effect of continued improvement in the crystal rectifier on the microwave receiver performance. The noise figures plotted are average values.

ability of the 3000-megacycle converter types made by the improved procedures and the older procedures is shown in Fig. 16.

The flexibility of the processes may be illustrated by comparison of two

very different units, the 1N26 and the 1N25. Though direct comparison of power handling ability is complicated by the fact that the burnout test methods employed in the development of the two codes were widely different, it may be stated conservatively that while the 1N26 would be damaged after absorbing something less than one watt peak pulse power, the 1N25 unit will withstand 25 watts peak or more. The 1N26 unit is, however, capable of satisfactory operation as a converter at a frequency of some 20 times that of the 1N25. These two units have been made by essentially the same procedures, the difference in properties being principally due to modification of alloy composition, heat treatment, and contact area.

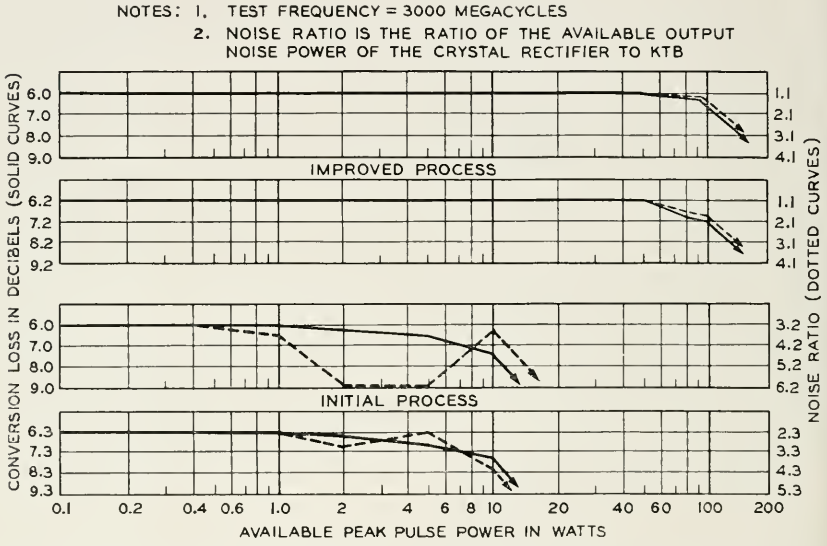


Fig. 16—Comparison of the radio frequency power handling ability of silicon crystal rectifiers prepared by different processes.

Prior to the process developments described above, in the interests of simplifying the field supply problem one general purpose unit, the type 1N21, had been made available for field use. However, it became obvious that the advantages of having but a single unit for field use could be retained only at a sacrifice in either power handling ability or high-frequency conversion loss. Since the higher power radar sets operated at the lower microwave frequencies, it seemed quite logical to employ the new processes to improve power handling ability at the lower microwave frequencies and to improve the loss and noise at the higher frequencies. A recommendation accordingly was made to the Services that different units be coded for operation at 3000 megacycles and at 10,000 megacycles. The decision in the matter was

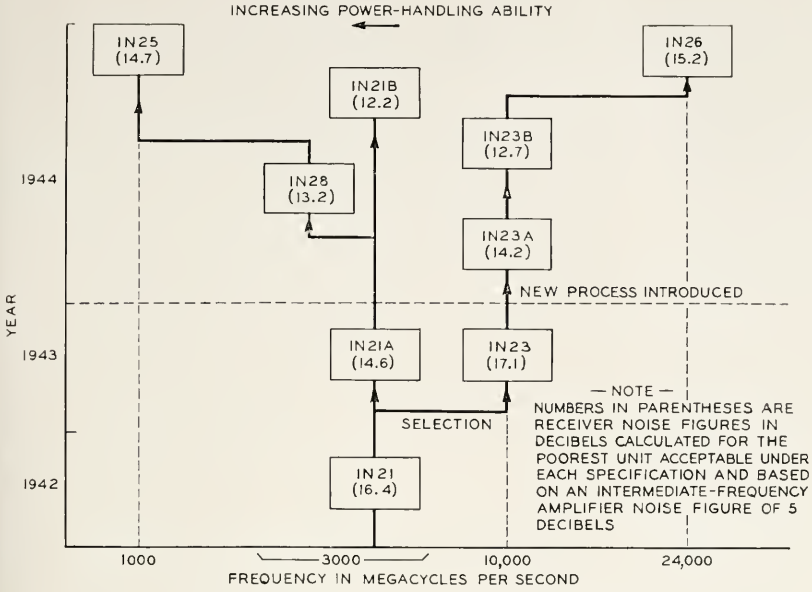


Fig. 17—Evolution of coded silicon crystal rectifiers.

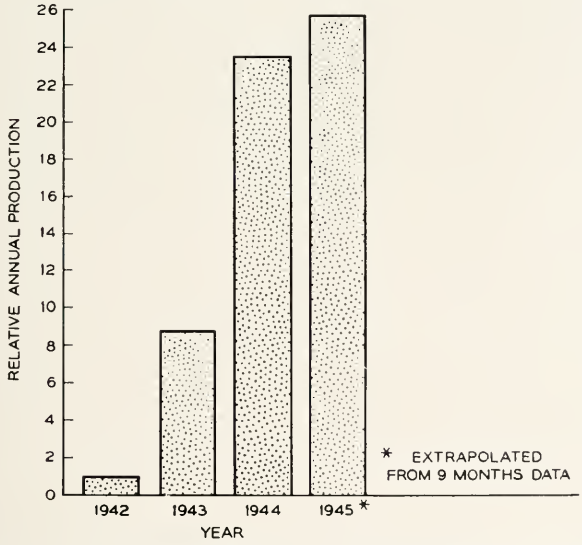


Fig. 18—Relative annual production of silicon crystal rectifiers at the Western Electric Company 1942–1945.

affirmative. The importance of this decision may be appreciated from the fact that it permitted the coding and manufacture of units such as the 1N21B and 1N28, high burnout units with improved performance at 3000

megacycles, and the 1N23B unit which was of such great importance in 10,000 megacycle radars because of its exceptionally good performance. From this stage in the development the diversification in types was quite rapid. The evolution of the coded units, of increasing power handling ability for a given performance level at a given frequency, and of better performance at a given frequency is graphically illustrated in Fig. 17. The large improvements in calculated receiver performance are again evident, especially when it is considered that the receiver performances given are for the poorest units which would pass the production test limits.

Extent of Manufacture and Utilization

An historical resumé of the development of crystal rectifiers would be incomplete if some description were not given of the extent of their manufacture and utilization. Commercial production of the rectifiers by Western Electric Company started in the early part of 1942 and through the war years increased very rapidly. Figure 18 shows the increase in annual production over that of the first year. By the latter part of 1944 the production rate was in excess of 50,000 units monthly. Production figures, however, reveal only a small part of the overall story of the development. The increase in production rate was achieved simultaneously with marked improvements in sensitivity, the improvements in process techniques being reflected in manufacture by the ability to deliver the higher performance units in increasing numbers.

The recent experience with the silicon rectifiers has demonstrated their utility as non-linear circuit elements at the microwave frequencies, that they may be engineered to exacting requirements of both a mechanical and electrical nature, and that they can be produced in large quantities. The deficiencies of the detector of World War I, which limited its utility and contributed to its retrogression, have now been largely eliminated. It is a reasonable expectation that the device will now find an extensive application in communications and other electrical equipment of a non-military character, at microwave as well as lower frequencies, where its sensitivity, low capacitance, freedom from aging effects, and its small size and low-power consumption may be employed advantageously.

Acknowledgements

The development of crystal rectifiers described in this paper required the cooperative effort of a number of the members of the staff of Bell Telephone Laboratories. The authors wish to acknowledge these contributions and in particular the contributions made by members of the Metallurgical group and the Holmdel Radio Laboratory with whom they were associated in the development.

End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator

By J. P. KINZER and I. G. WILSON

Formulas are given for the calculation of the current streamlines and intensity in the walls of a circular cylindrical cavity resonator. Tables are given which permit the calculation to be carried out for many of the lower order modes.

The integration of $\int_0^x \frac{J_\ell(x)}{J'_\ell(x)}\,dx$ is discussed; the integration is carried out for $\ell = 1$, 2 and 3 and tables of the function are given.

The current distribution for a number of modes is shown by plates and figures.

Introduction

In waveguides or in cavity resonators, a knowledge of the electromagnetic field distribution is of prime importance to the designer. Representations of these fields for the lower modes in rectangular, circular and elliptical waveguide, as well as coaxial transmission line, have frequently been described.

For the most part, however, these representations have been diagrammatic or schematic, intended only to give a general physical picture of the fields. In actual designs, such as high Q cavities for use as echo boxes,[1] accurately made plates of the distributions were found necessary to handle adequately problems of excitation of the various modes and of mode suppression.

One use of the charts is to determine where an exciting loop or orifice should be located and how the field should be oriented for maximum coupling to a particular mode. Optimum locations for both launchers and absorbers can be found. Naturally, when attention is concentrated on a single mode these will be located at the maximum current density points. If, however, two or more modes can coexist, and only one is desired, compromise locations can sometimes be found which minimize the unwanted phenomena.

Also, in a cylindrical cavity resonator of high Q with diameter large compared with the operating wavelength, there are many high order modes of oscillation whose resonances fall within the design frequency band. Some of these are undesired and one of the objectives of a practical design is to reduce their responses to a tolerable amount. This process is termed

[1] "High Q Resonant Cavities for Microwave Testing," Wilson, Schramm, Kinzer, *B.S.T.J.*, July 1946.

"suppression of the extraneous modes". In this process, an exact knowledge of the distribution of the currents in the cavity walls has been found highly useful.

For example, it has been found experimentally that annular cuts in the end plates of the cylinder give a considerable amount of suppression to many types of extraneous modes with very little effect on the performance of the desired *TE* 01n mode. These cuts are narrow slits concentric with the axis of the cylinder and going all the way through the metallic end plates into a dielectric beyond.[2] The physical explanation is that an annular slit cuts through the lines of current flow of the extraneous modes, and thereby interrupts the radial component of current and introduces an impedance which damps, or suppresses, the mode. For the *TE* 01n mode, the slits

	TE Modes	*TM* Modes
End Plates	$H_\rho = \frac{k_3}{k_1} J_\ell'(k_1\rho)\cos\ell\theta$ $H_\theta = -\ell\frac{k_3}{k_1}\frac{J_\ell(k_1\rho)}{k_1\rho}\sin\ell\theta$	$H_\rho = \ell\frac{J_\ell(k_1\rho)}{k_1\rho}\sin\ell\theta$ $H_\theta = J_\ell'(k_1\rho)\cos\ell\theta$
Side Walls	$H_\theta = \left[-\ell\frac{k_3}{k_1}\frac{J_\ell(k_1 D/2)}{k_1 D/2}\right][\sin\ell\theta\cos k_3 z]$ $H_z = J_\ell(k_1 D/2)\cos\ell\theta\sin k_3 z$	$H_\theta = J_\ell'(k_1 D/2)\cos\ell\theta\cos k_3 z$ $H_z = 0$

$$k = \frac{2\pi}{\lambda} = k_1^2 + k_3^2$$

$$k_1 = \frac{2r}{D} \qquad k_3 = \frac{n\pi}{L}$$

r = m^{th} root of $J_\ell(x) = 0$ for *TM* Modes.
= m^{th} root of $J_\ell'(x) = 0$ for *TE* Modes.

D = cavity diameter

L = cavity length

Fig. 1—Components of H vector at walls of circular cylinder cavity resonator.

are parallel to the current streamlines and there is no such interruption; presumably there is a slight increase in current density alongside the slit,

[2] Similar cuts through the side wall of the cylinder in planes perpendicular to the cylinder axis are also beneficial, but are more troublesome mechanically.

as the current formerly on the surface of the removed metal crowds over onto the adjacent metal, but this is a second-order effect.

To determine the best location of such cuts, therefore, it is necessary to know the vector distributions of the wall currents for the various modes. This current vector, I, is proportional to and perpendicular to the magnetic vector, H, of the field at the surface. Expressions for the components of the H-vector at the surfaces of the end plates and side walls are given in Fig. 1.

END PLATE: *Contour Lines*

At the end plates, the magnitude of the H-vector at any point is given by:

$$H^2 = H_\rho^2 + H_\theta^2. \tag{1}$$

Now substitute values of H_ρ and H_θ from Fig. 1 into (1); drop any constant factors common to H_ρ and H_θ as these can be swallowed in a final proportionality constant; introduce the new variable x:

$$x = k_1 \rho = r \frac{\rho}{R}. \tag{2}$$

where $R = D/2$ = cavity radius. Thus is obtained:

$$H^2 = [J'_\ell(x) \cos \ell\theta]^2 + \left[\frac{\ell}{x} J_\ell(x) \sin \ell\theta\right]^2. \tag{3}$$

Now J_ℓ and J'_ℓ, are expressed in terms of $J_{\ell-1}$ and $J_{\ell+1}$ and a further reduction leads to:

$$H^2 = (J_{\ell-} \cos \ell\theta)^2 + (J_{\ell+} \sin \ell\theta)^2 \tag{4}$$

where

$$J_{\ell-} = J_{\ell-1}(x) - J_{\ell+1}(x) \tag{5}$$

and

$$J_{\ell+} = J_{\ell-1}(x) + J_{\ell+1}(x) \tag{6}$$

The formulas (4) to (6) apply to both TE and TM modes. The values obtained depend on r, which is different for each mode.

When $\theta = 0$, I is proportional to $J_{\ell-}$ and when $\theta = \pi/2\ell$, I is proportional to $J_{\ell+}$. Relative values of I are thus easily calculated for these cases, once tables of J_ℓ are available. Such tables have been prepared and are attached. For TE modes, when $\theta = 0$, $H_\theta = 0$, and the currents are all in the θ direction. For TM modes, when $\theta = 0$, $H_\rho = 0$, and the currents are all in the ρ-direction. When $\theta = \pi/2\ell$, the converse holds.

Figures 3 to 18 are a set of curves showing the relative magnitude of H (or I) for several of the lower order TE and TM modes. The abscissae

are relative radius, i.e., ρ/R; the ordinates are relative magnitude referred to the maximum value. The drawings also give $r = \pi D/\lambda_c$ for each mode, where λ_c is the cutoff wavelength in a circular guide of diameter D. Values for any point of the surface of the end plate can be calculated by using these curves in conjunction with equation (4).

In general, for each mode there are certain radii at which the current flow is entirely radial, ($I_\theta = 0$). At these radii, which correspond to zeros of $J_\ell(x)$ or $J'_\ell(x)$, the annular cuts mentioned in the introduction are quite effective. However, the maxima of I_ρ do not coincide with the zeros of I_θ; and a more sophisticated treatment gives the best radius as that which maximizes ρI_ρ^2. Values of the relative radius for this last condition are given in Table IV.

Contour lines of equal relative current intensity are obtained by setting H^2 constant in (4), which then expresses a relation between x and θ. The easiest and quickest way to solve (4) is graphically, by plotting H vs. x for different values of θ.

END PLATE: *Current Streamlines*

It is easy to show that the equations of the current streamlines are given by the solutions of the differential equation

$$\frac{d\rho}{d\theta} = -\rho \frac{H_\theta}{H_\rho}. \tag{7}$$

In the case of the TE modes, (7) is easily solved by separation of the variables, leading to the final result:

$$J_\ell(x) \cos \ell\theta = C \tag{8}$$

in which C is a parameter whose value depends on the streamline under consideration. In the TE modes, the E-lines in the interior of the cavity also satisfy (8), hence a plot of the current streamlines in the end plate serves also as a plot of the E lines.

In the case of the TM modes, (7) is not so easily solved. Separation of the variables leads to:

$$-\log \sin \ell\theta = \int \frac{\ell^2 J_\ell(x)}{x^2 J'_\ell(x)}\, dx. \tag{9}$$

The right-hand side of (9) can be reduced somewhat, yielding

$$-\log \sin \ell\theta = \log [xJ'_\ell(x)] + \int \frac{J_\ell(x)}{J'_\ell(x)}\, dx \tag{10}$$

but no further reduction is possible. The remaining integral represents a new function which must be tabulated. Its evaluation is discussed at

length in the Appendix, where it is denoted by $F_\ell(x)$. Table II of the Appendix gives its values (for $\ell = 1, 2$ and 3) and also those of $G_\ell(x)$ where

$$F_\ell(x) = -\log G_\ell(x) \tag{11}$$

Thus (10) becomes

$$-\log \sin \ell\theta = \log [x J'_\ell(x)/G_\ell(x)] + C' \tag{12}$$

and the final equation for the current streamlines is

$$[xJ'_\ell(x)/G_\ell(x)] \sin \ell\theta = C \tag{13}$$

where C is a parameter as before.

It is not difficult to show that $G_\ell(x)/J'_\ell(x)$ has zeros at the zeros of $J'_\ell(x)$. For these values of x, $\sin \ell\theta = 0$ whatever the value of C, and all streamlines converge on (or diverge from) $2\ell m$ points on the end plate.

The flow lines of (13) are orthogonal to the family (8) and could readily be drawn in this manner. However, better accuracy is obtained by plotting (13).

END PLATE: *Distributions*

The 32 attached plates show the distribution of current in the end plates of a circular cylinder cavity resonator for a number of modes.

In the first set of 21, the scaling is such that the diameters of the figures are proportional to those of circular waveguides which would have the same cutoff frequency. This group is of particular interest to the waveguide engineer.

In a second group of 11, the scaling is such as to make the outside diameters of the cylinders uniform. This group is of particular interest to a cavity designer.

This distribution is a vector function of position; that is, at each point in the end plate the surface current has a different direction of flow and a different magnitude or intensity. The variation in current intensity is represented by ten degrees of background shading. The lightest indicates regions of least current intensity and the darkest greatest intensity. The direction of current flow is shown by streamlines. Streamlines are lines such that a tangent at any point indicates the direction of current flow at that point.

The modes represented are the

TE 01, 02, 03	*TM* 01, 02, 03
TE 11, 12, 13	*TM* 11, 12, 13
TE 21, 22, 23	*TM* 21, 22
TE 31, 32	*TM* 31, 32

in the nomenclature which has become virtually standard. In this system, *TE* denotes transverse electric modes, or modes whose electric lines lie in planes perpendicular to the cylinder axis; *TM* denotes transverse magnetic modes, or modes whose magnetic lines lie in transverse planes. The first numerical index refers to the number of nodal diameters, or to the order of the Bessel function associated with the mode. The second numerical index refers to the number of nodal circles (counting the resonator boundary as one such) or to the ordinal number of a root of the Bessel function associated with the mode. On the end plates, the distribution does not depend upon the third index (number of half wavelengths along the axis of the cylinder) used in the identification of resonant modes in a cylinder. This considerably simplifies the problem of presentation. The orientation of the field inside the cavity and hence the currents in the end plate depend on other things; thus the orientation of the figures is to be considered arbitrary.

The plates also apply to the corresponding modes of propagation in a circular waveguide as follows: The background shading represents the instantaneous relative distribution of energy across a cross section of guide. For *TE* modes, the current streamlines depict the *E* lines; for the *TM* modes, they depict the projection of the *E* lines on a plane perpendicular to the cylinder axis.

Side Wall:

The current distribution in the side walls is easily obtained from the field equations of Fig. 1. For *TM* modes, the currents are entirely longitudinal; their magnitudes vary as $\cos \ell\theta \cos n\pi z/L$. This distribution is so simple as not to require plotting.

For *TE* modes, the situation is more complicated, since both H_z and H_θ exist along the side wall. The current streamlines are given by the solutions of the differential equation

$$\frac{dz}{d\theta} = -\frac{DH_\theta}{2H_z}. \tag{14}$$

By separation of the variables, the solution is found to be

$$\log (C \cos \ell\theta) = \left[\frac{k_1}{k_3}\right]^2 \log \cos k_3 z. \tag{15}$$

Contour lines of constant magnitude of the current are given by

$$\left(\frac{2k_3 \ell}{k_1^2 D} \sin \ell\theta \cos k_3 z\right)^2 + (\cos \ell\theta \sin k_3 z)^2 = K^2. \tag{16}$$

In the above, C and K are parameters, different values of which correspond to different streamlines or contour lines, respectively.

Since both streamlines and contours are periodic in z and θ, it is not essential to represent more than is covered in a rectangular piece of the side wall corresponding to quarter periods in z and θ. These are covered in a length $\frac{L}{2n}$ along the cavity and in a distance $\frac{\pi D}{4\ell}$ around the cavity. If such a piece of the surface be rolled out onto a plane it forms a rectangle of proportions $\frac{\pi n D}{2\ell L}$.

The difficulty in depicting the side wall currents of TE modes, as compared with the end plate currents, is now apparent. For the end plate, the "proportions" are fixed as being a circle. Furthermore, for a given ℓ, as m increases the effect is merely to add on additional rings to the previous streamline and contour plots. Here, however, the proportions of the rectangle are variable, in the first place. And for a given rectangle the streamlines and contours both change as ℓ and m are varied. Another way of expressing the same idea is that for end plates the current distribution does not depend upon the mode index n, and varies only in an additive way with the index m, whereas for the side walls the distribution depends in nearly equal strength on ℓ, m and n.

Some simplification of the situation is accomplished by introducing two new parameters, the "shape" and the "mode" parameters, defined by:

$$S = \frac{\pi n D}{2\ell L} \qquad M = \frac{\ell}{r} \tag{17}$$

and two new variables

$$Z = k_3 z \qquad \phi = \ell\theta. \tag{18}$$

Substitution of the above, and also the expressions for k_1 and k_3 (see Fig. 1) into (15) and (16) yields

$$\cos Z = C(\cos\phi)^{S^2 M^2} \quad \text{(streamlines)} \tag{19}$$

$$\cos Z = \left[\frac{K^2 - \cos^2\phi}{(S^2 M^4 \sin^2\phi - \cos^2\phi)}\right]^{1/2} \quad \text{(contours)}. \tag{20}$$

For given proportions S, one can calculate the streamlines and contours for various values of M. Thus a "square array" of side wall currents can be prepared, such as shown on Fig. 2.

The mode parameter, M, in the physical case takes on discrete values which depend on the mode. Some of its values are given in the following table. They all lie between 0 and 1 and there are an infinite number of them.

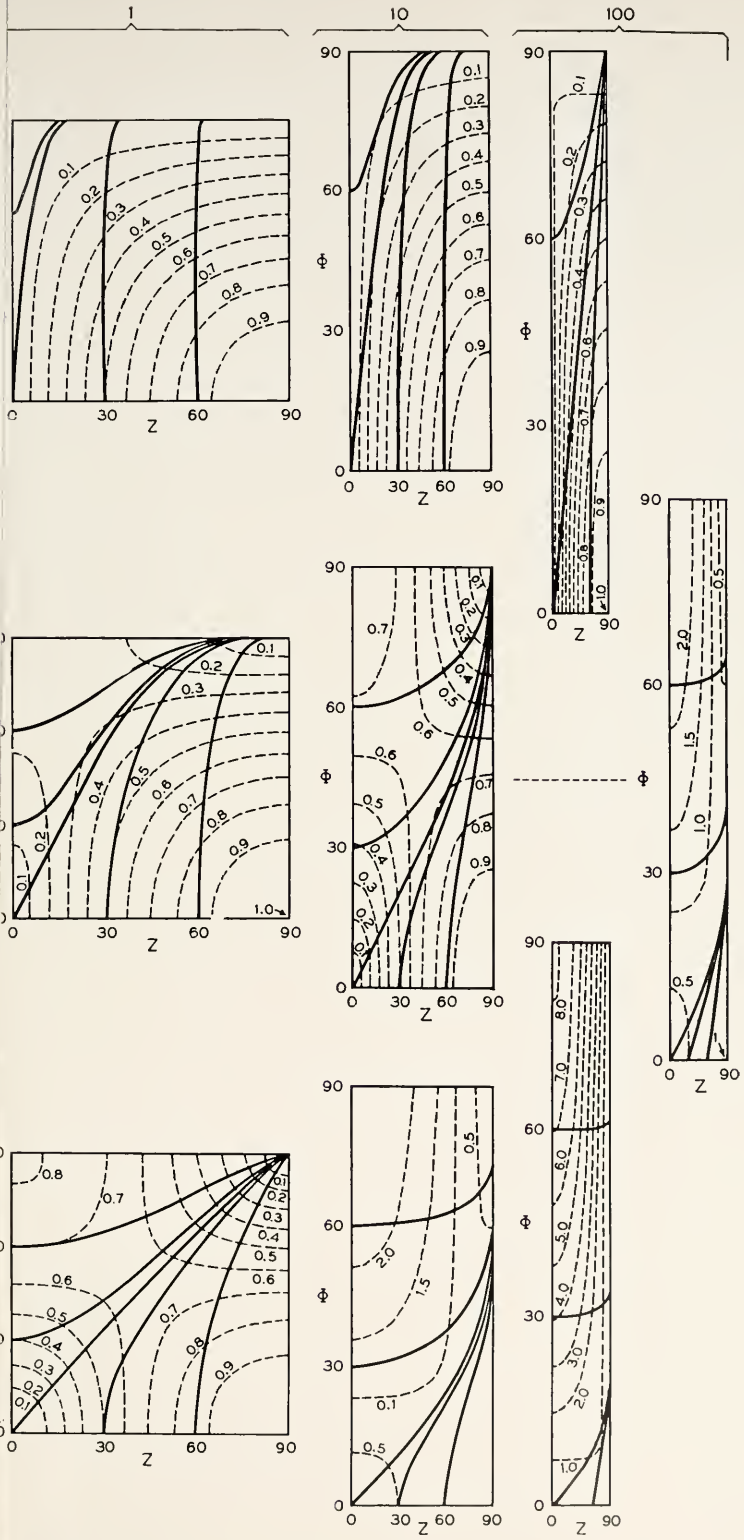
VALUE OF $M = \ell/r$ FOR TE MODES

ℓ	1	2	3	4	5	6	10	15	20
$m = 1$	.5432	.6549	.7141	.7522	.7793	.8000	.8495	.8813	.9001
2	.1875	.2982	.3743	.4309	.4753	.5113	.6080	.6774	
3	.1172	.2006	.2644	.3154	.3575	.3930	.4945	.5730	
4	.0854	.1519	.2057	.2506	.2888	.3219	.4209		

For any given mode in any given cavity, the values of S and M can be calculated from (17). In general, these values will not coincide with those which have been plotted, but by the same token, they will lie among a group of four combinations which have been plotted. Since the changes in distribution are smooth, mental two-way interpolation will present no difficulty.

ACKNOWLEDGMENT

The final plates depicting the current distributions are the result of the efforts of many individuals in plotting, spray tinting of the background, inking of the streamlines on celluloid overlay, and photographing. Special mention must be made, however, of the contribution of Miss Florence C. Larkey, who carried out all the lengthy calculations of the tables hereto attached and of the necessary data for the plotting.



or TE modes ($l > 0$).

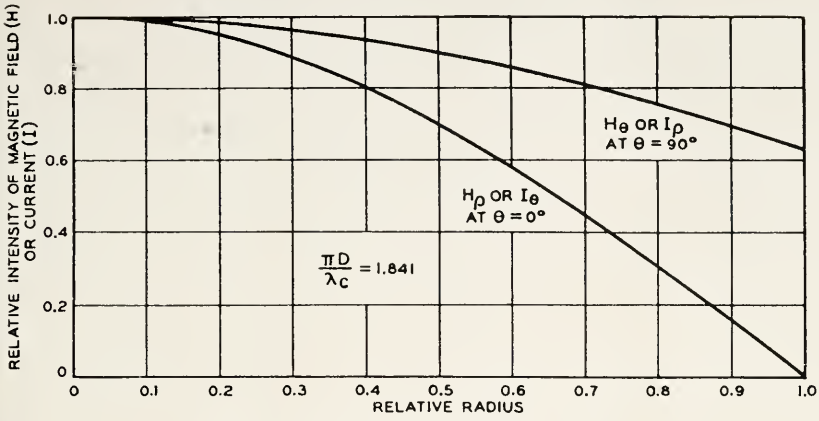


Fig. 3—End plate currents in TE 11 mode.

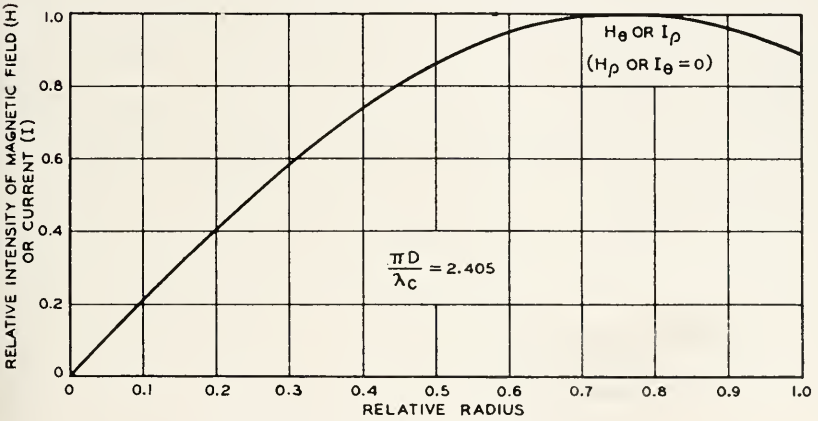


Fig. 4—End plate currents in TM 01 mode.

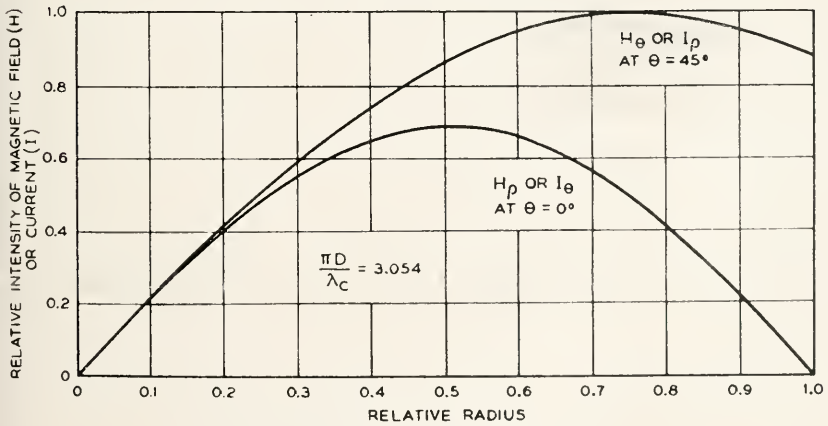


Fig. 5—End plate currents in TE 21 mode.

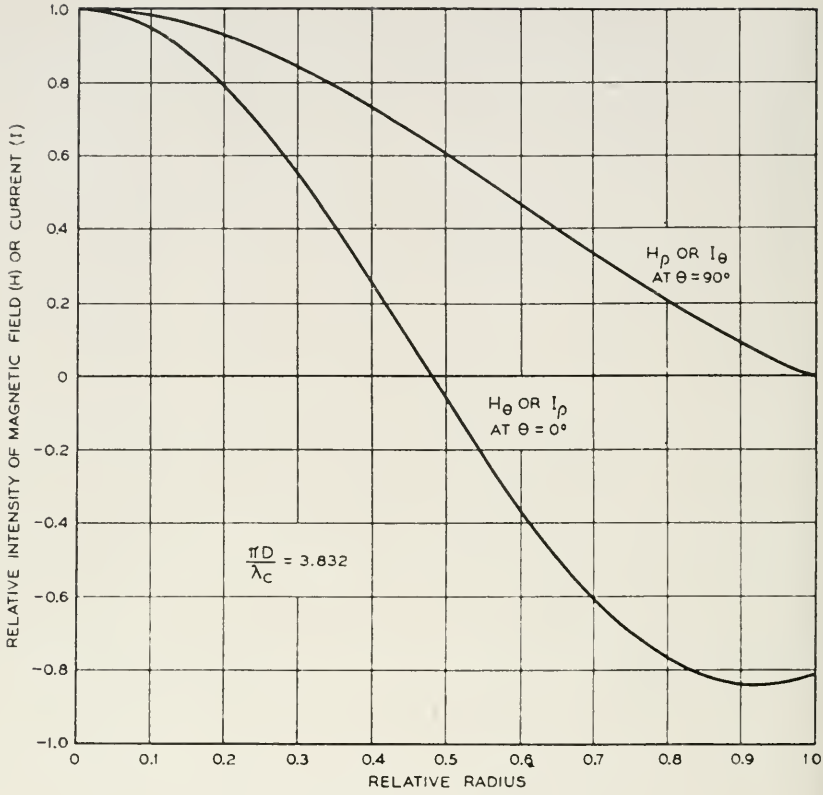


Fig. 6—End plate currents in TM 11 mode.

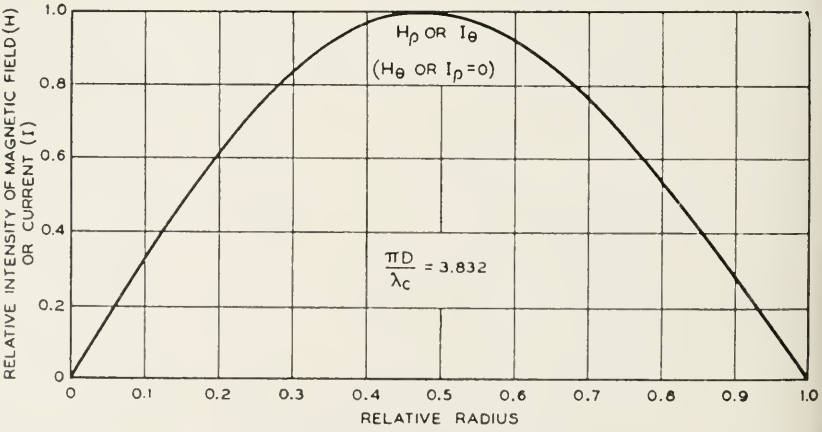


Fig. 7—End plate currents in TE 01 mode.

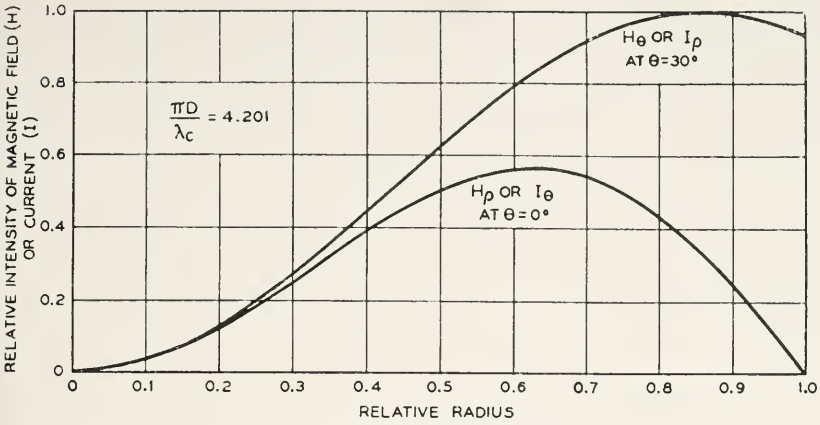


Fig. 8—End plate currents in TE 31 mode.

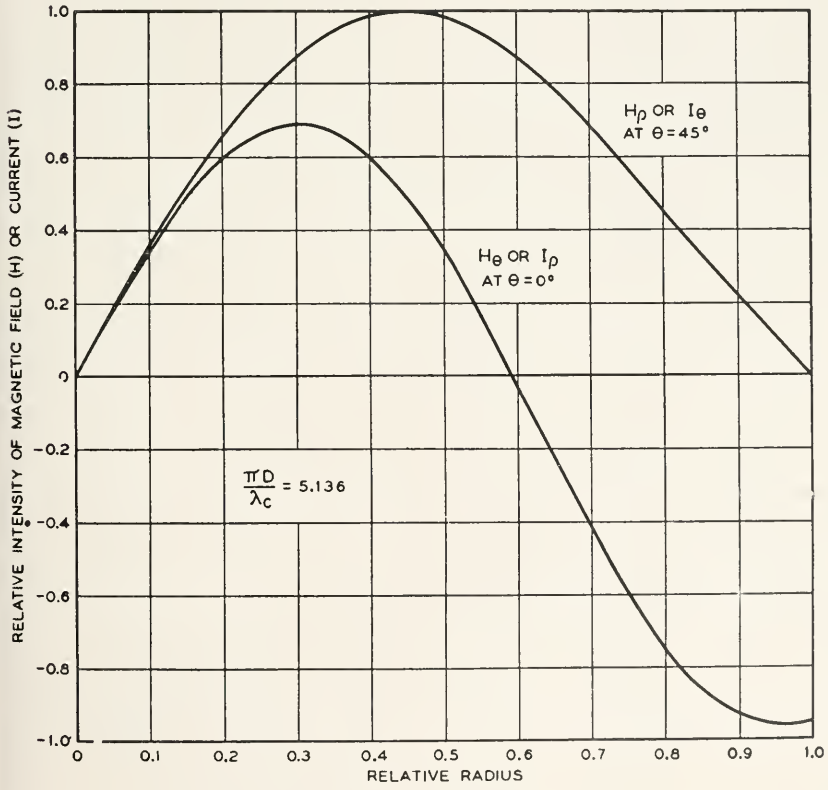


Fig. 9—End plate currents in TM 21 mode.

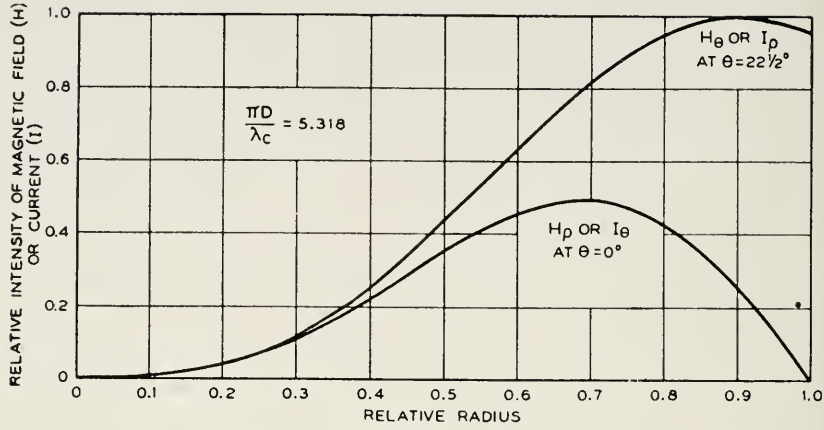


Fig. 10—End plate currents in TE 41 mode.

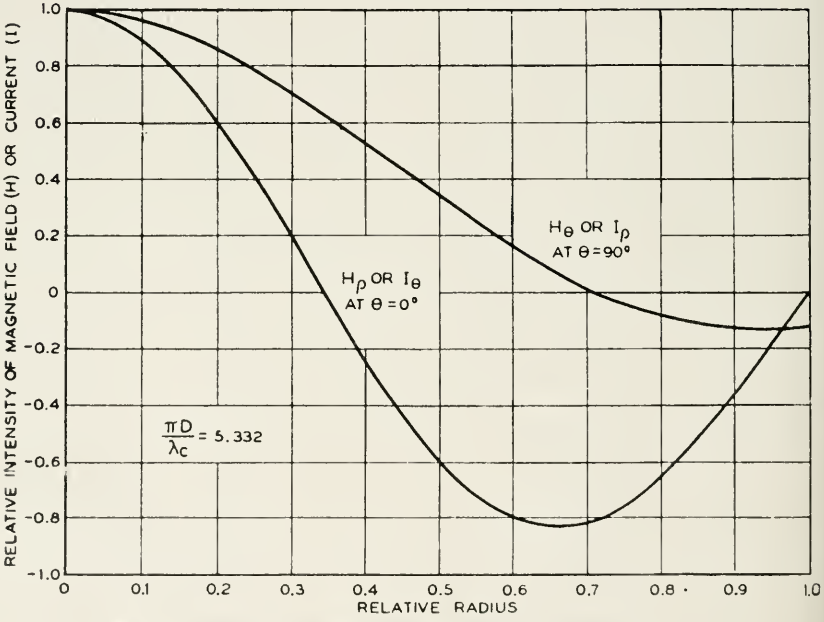


Fig. 11—End plate currents in TE 12 mode.

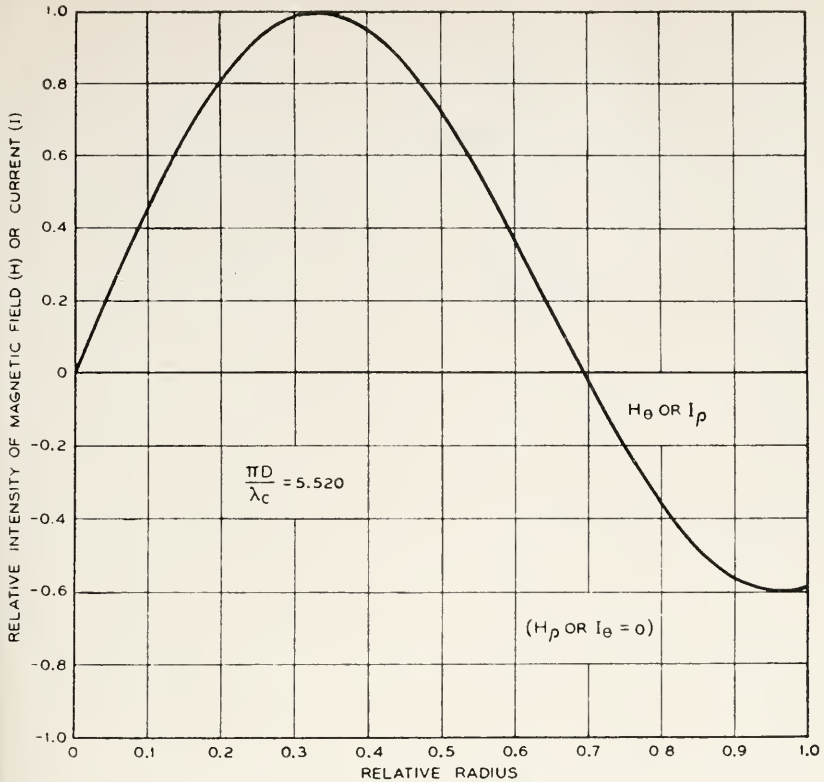


Fig. 12—End plate currents in TM 02 mode.

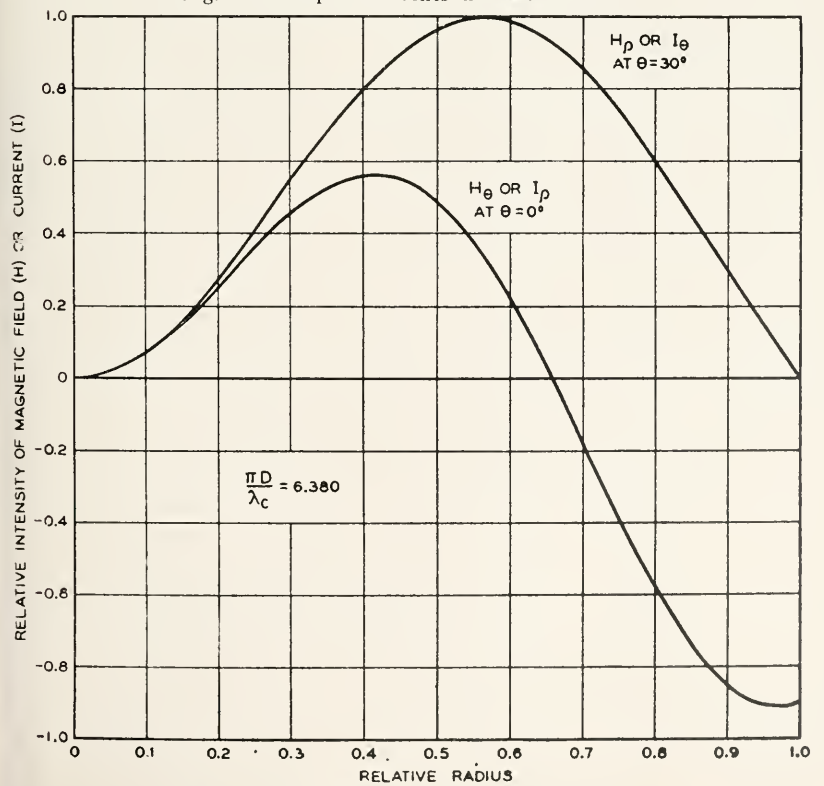


Fig. 13—End plate currents in TM 31 mode.

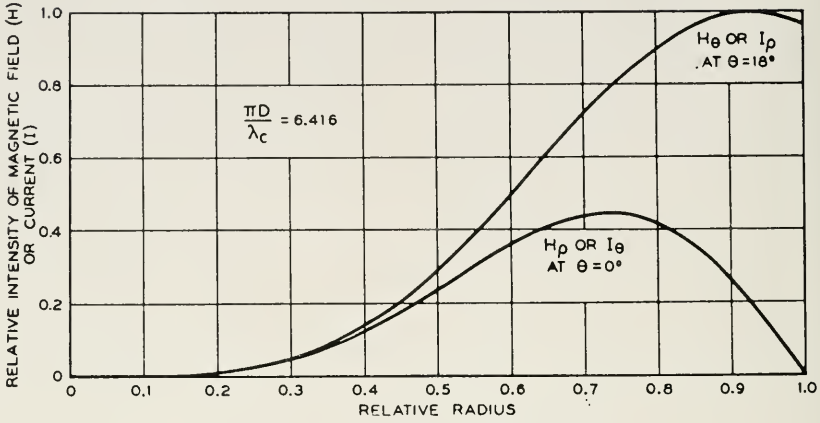


Fig. 14—End plate currents in TE 51 mode.

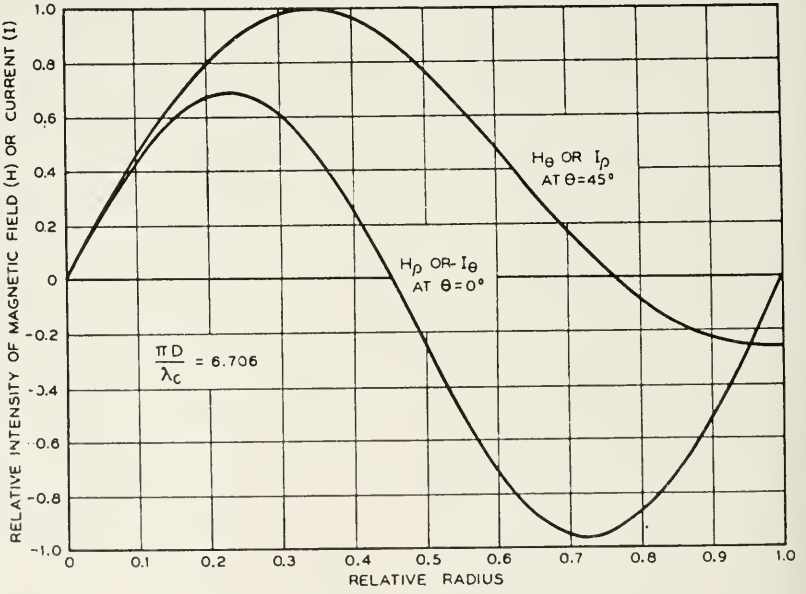


Fig. 15—End plate currents in TE 22 mode.

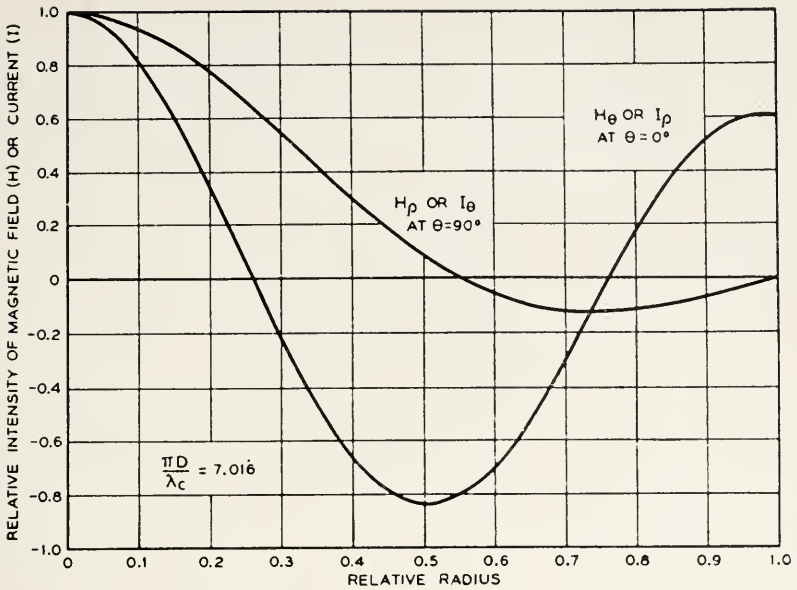


Fig. 16—End plate currents in TM 12 mode.

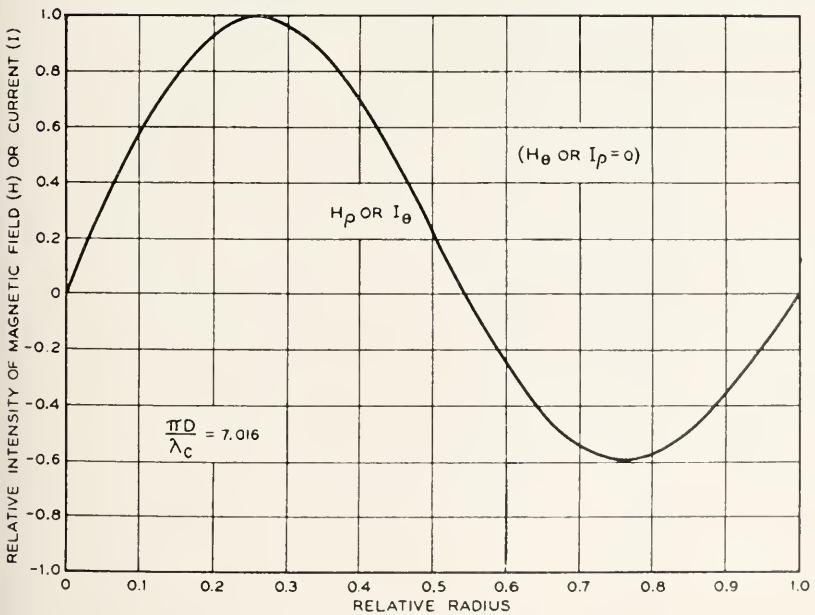


Fig. 17—End plate currents in TE 02 mode.

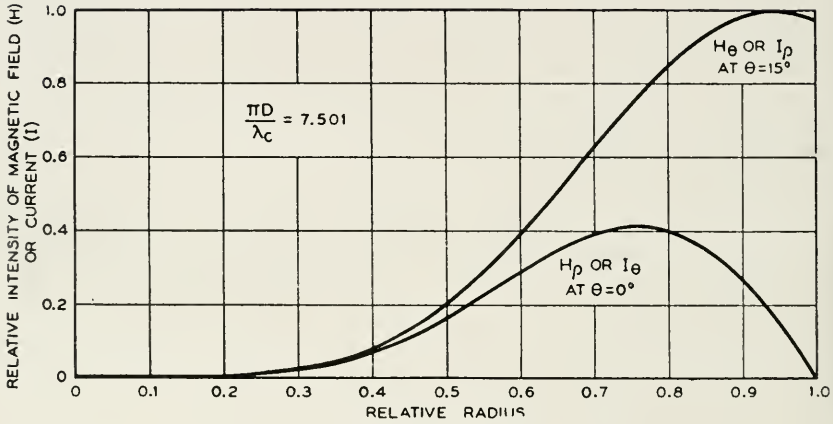


Fig. 18—End plate currents in TE 61 mode.

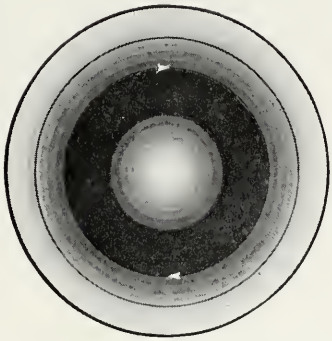


Fig. 19—TE 01 mode.

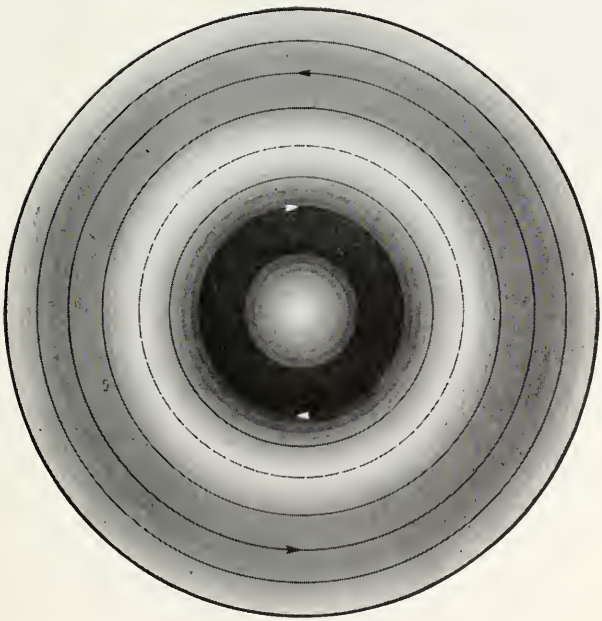


Fig. 20—TE 02 mode.

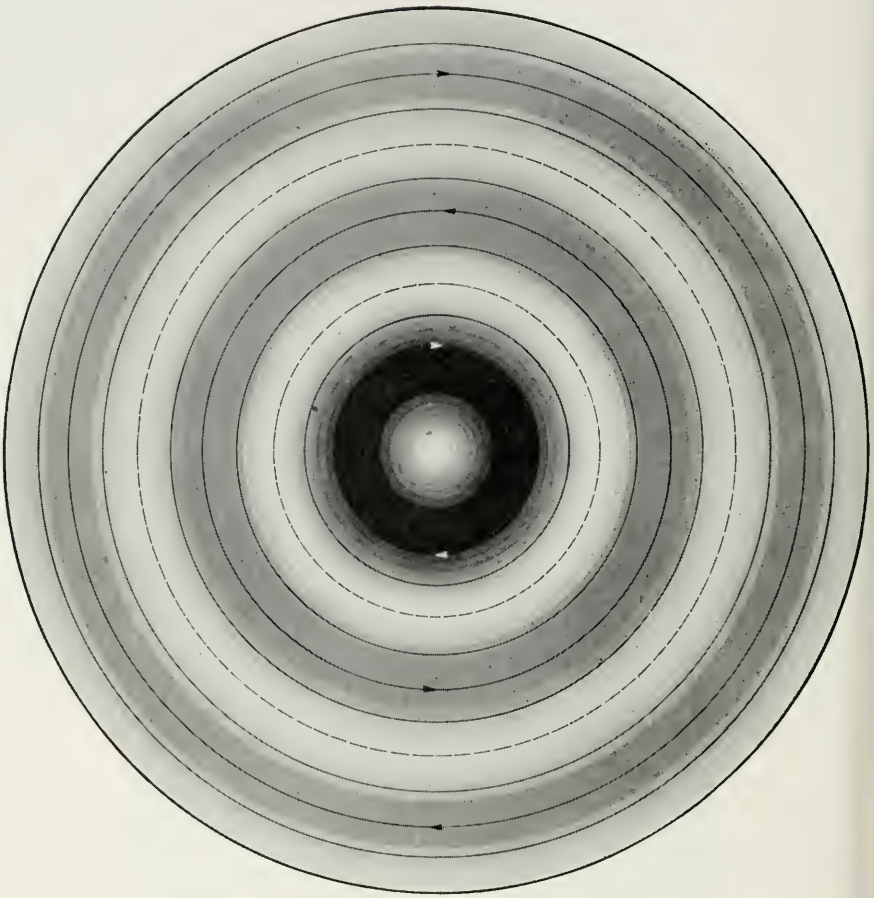


Fig. 21—TE 03 mode.

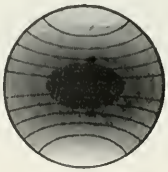


Fig. 22—TE 11 mode.

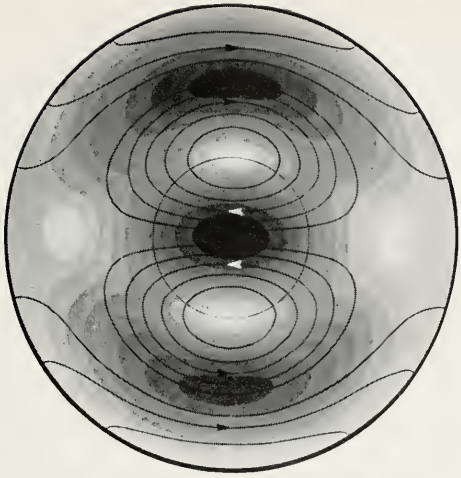


Fig. 23—TE 12 mode.

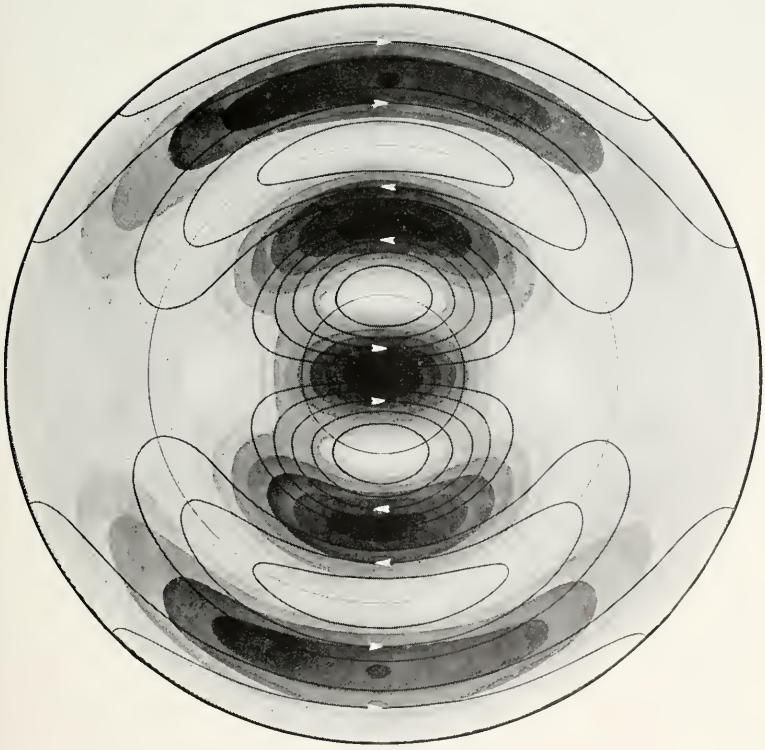


Fig. 24—TE 13 mode.

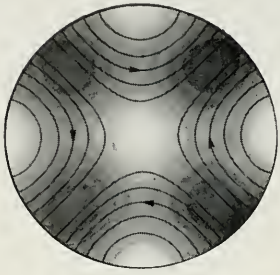


Fig. 25—TE 21 mode.

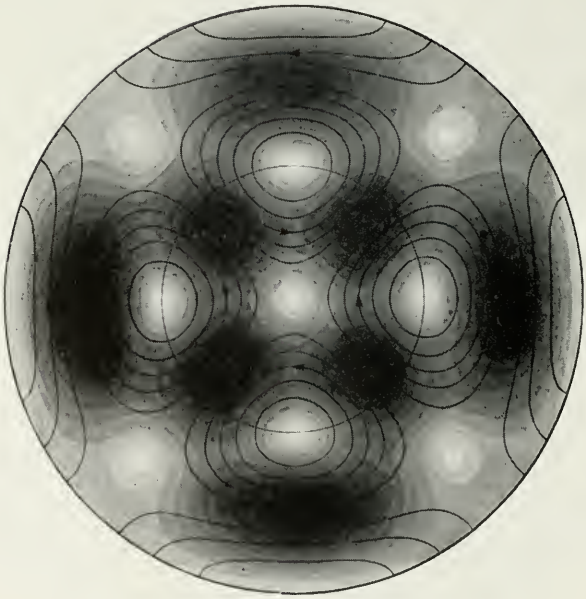


Fig. 26—TE 22 mode.

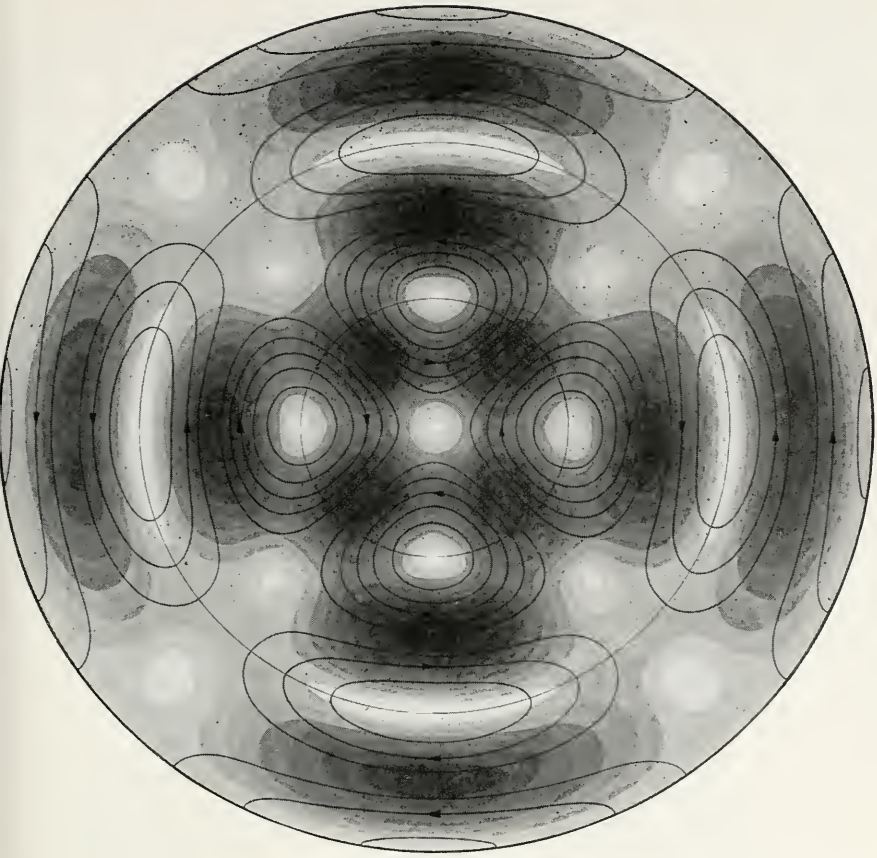


Fig. 27—TE 23 mode.

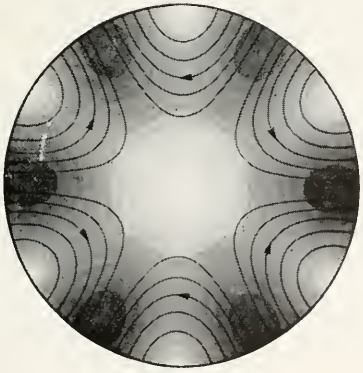


Fig. 28—TE 31 mode.

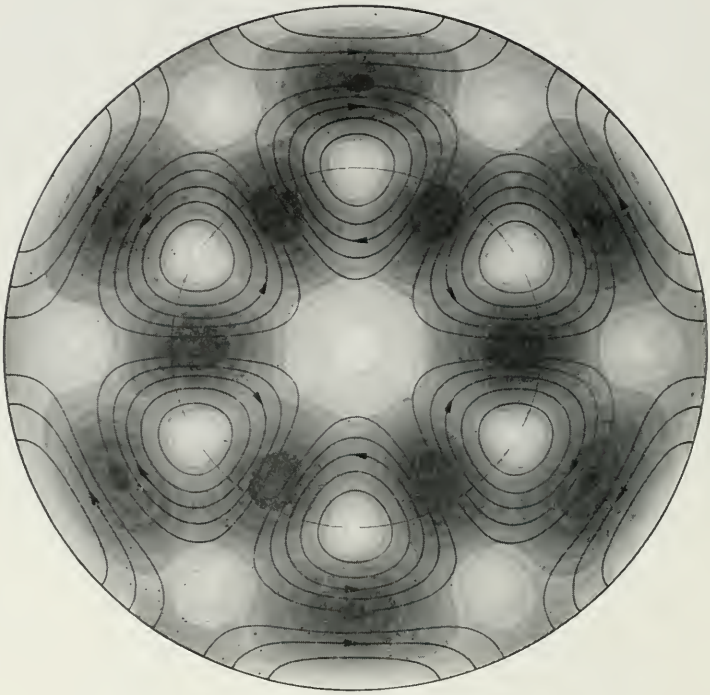


Fig. 29—TE 32 mode.

Fig. 30—TM 01 mode.

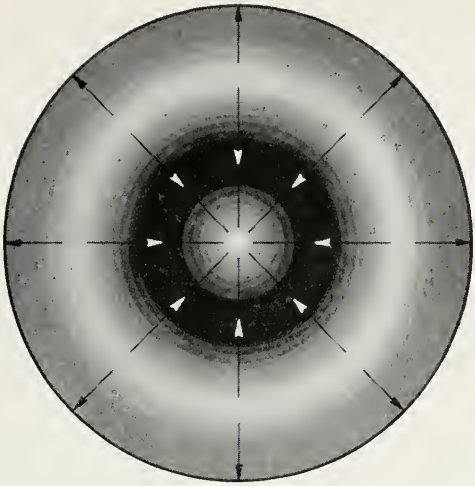


Fig. 31—TM 02 mode.

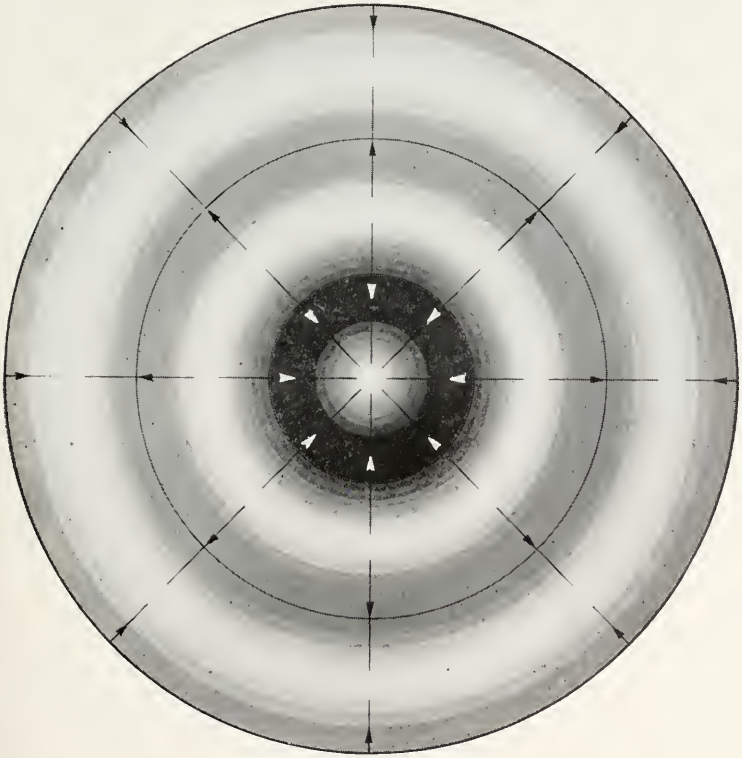


Fig. 32—TM 03 mode.

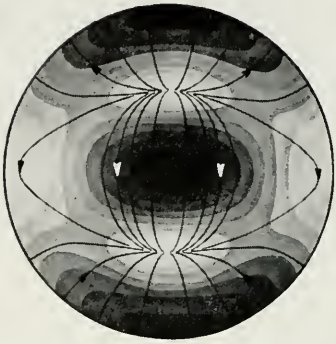


Fig. 33—TM 11 mode.

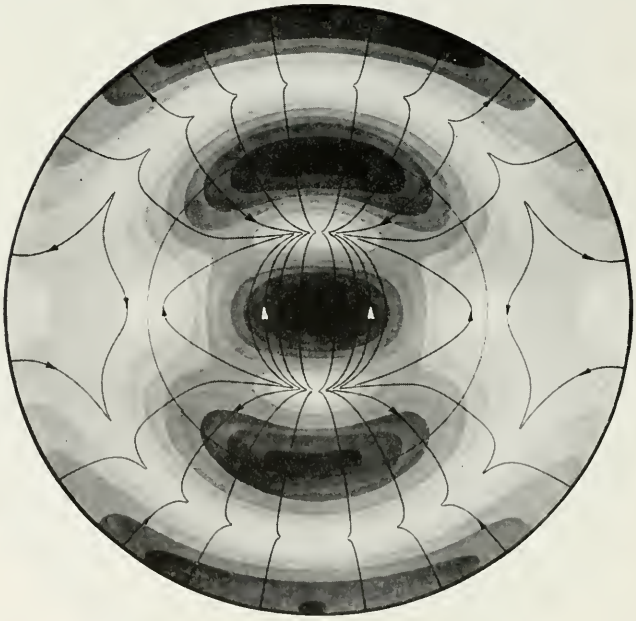


Fig. 34—TM 12 mode.

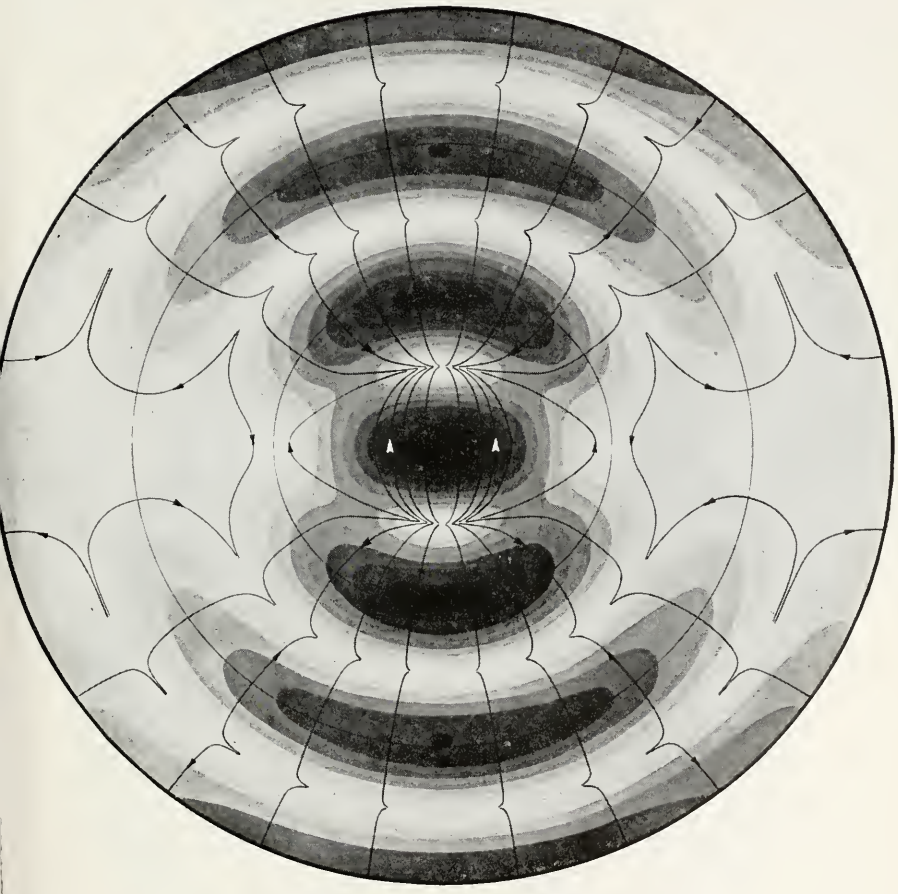


Fig. 35—TM 13 mode.

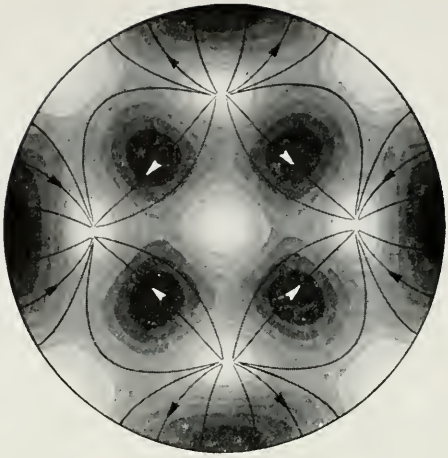


Fig. 36—TM 21 mode.

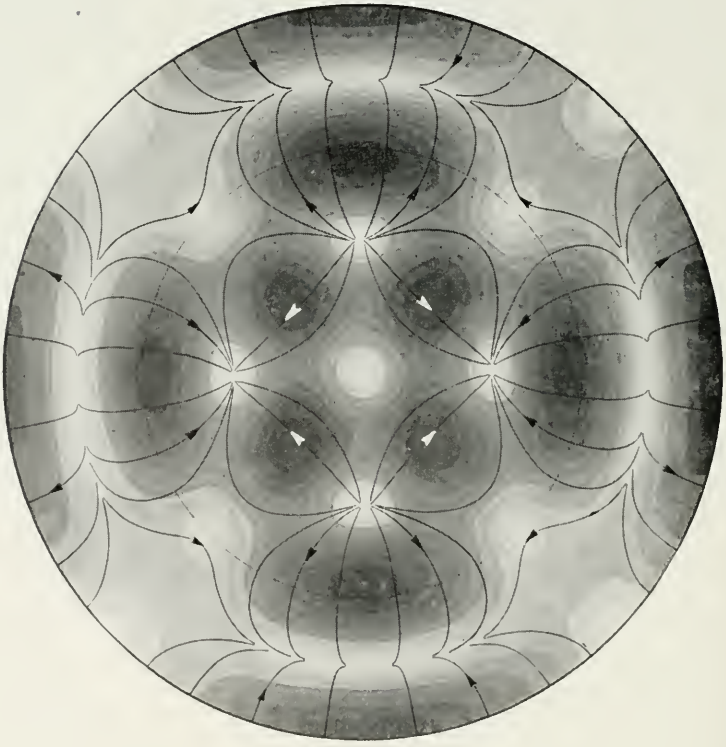


Fig. 37—TM 22 mode.

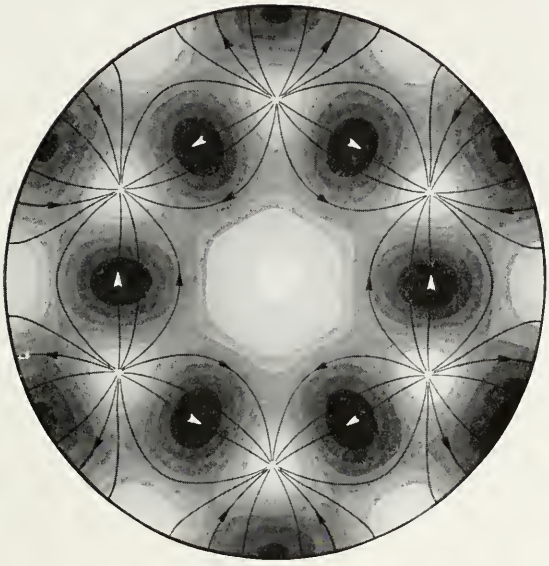


Fig. 38—TM 31 mode.

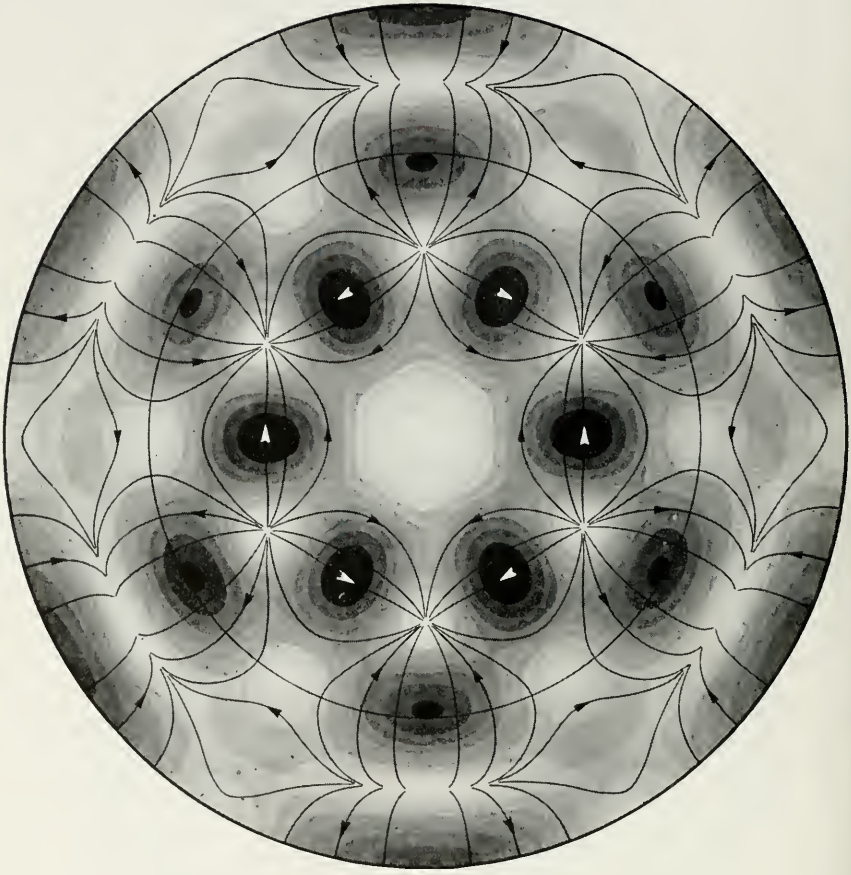


Fig. 39—TM 32 mode.

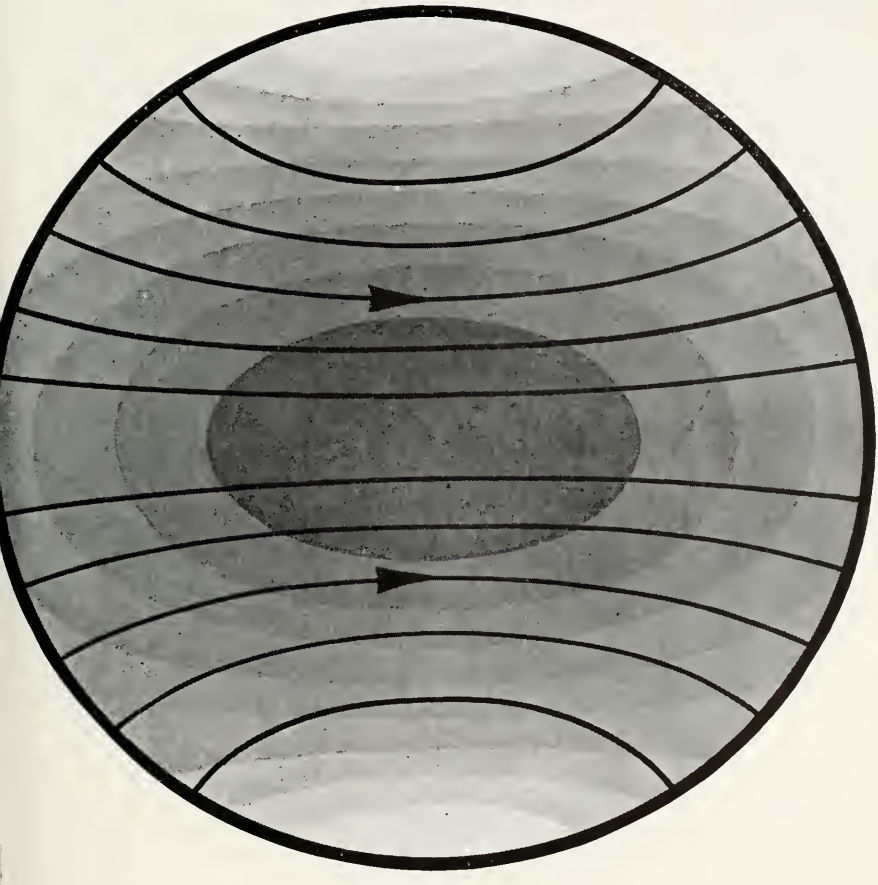


Fig. 40—TE 11 mode.

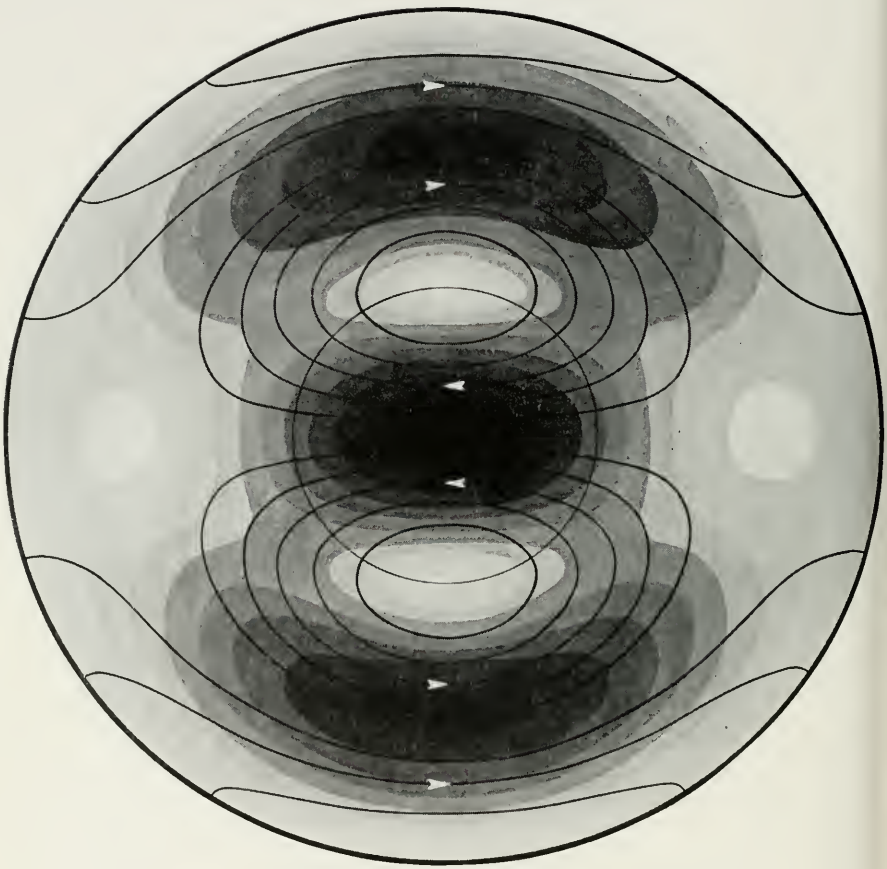


Fig. 41—TE 12 mode.

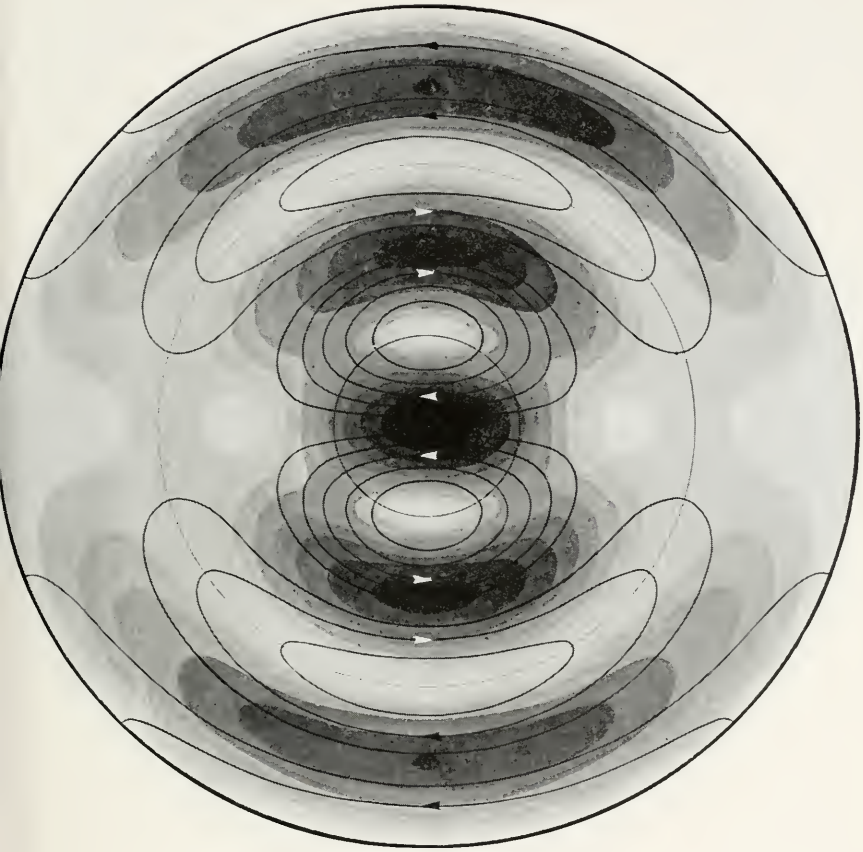


Fig. 42—TE 13 mode.

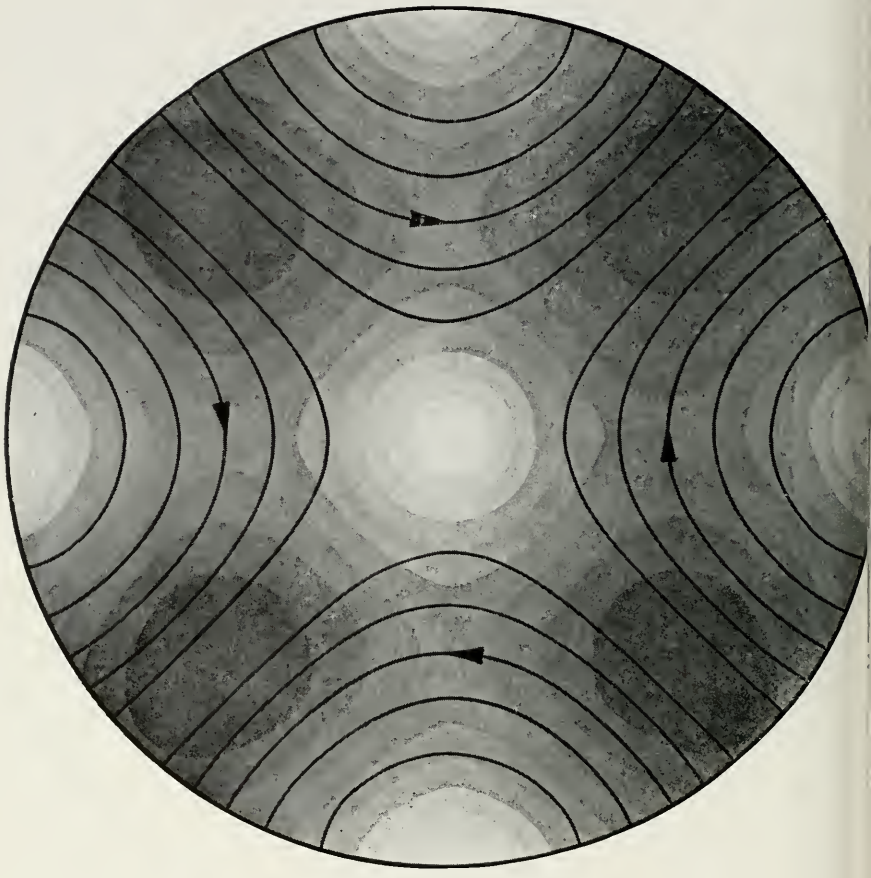


Fig. 43—TE 21 mode.

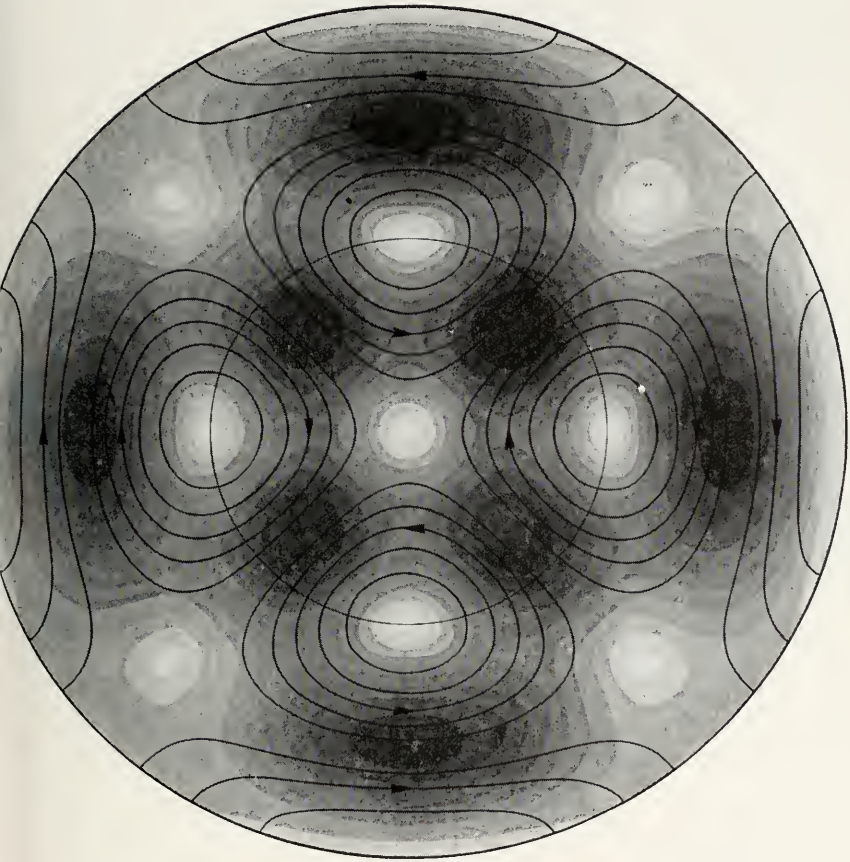


Fig. 44—TE 22 mode.

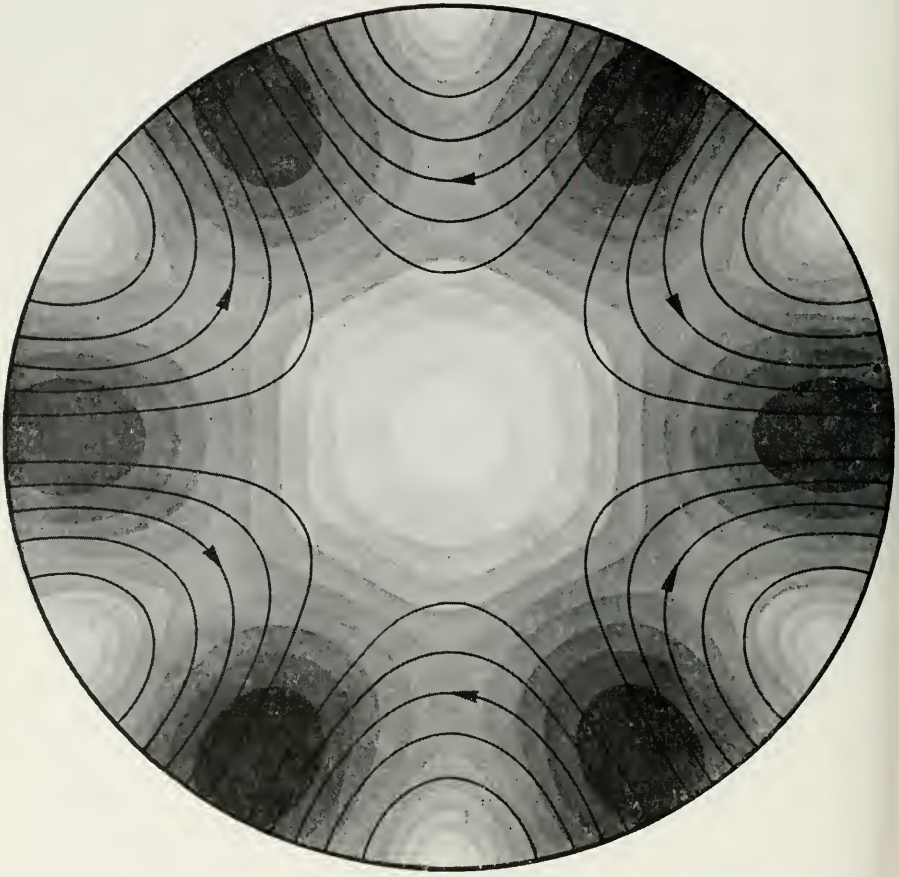


Fig. 45—TE 31 mode.

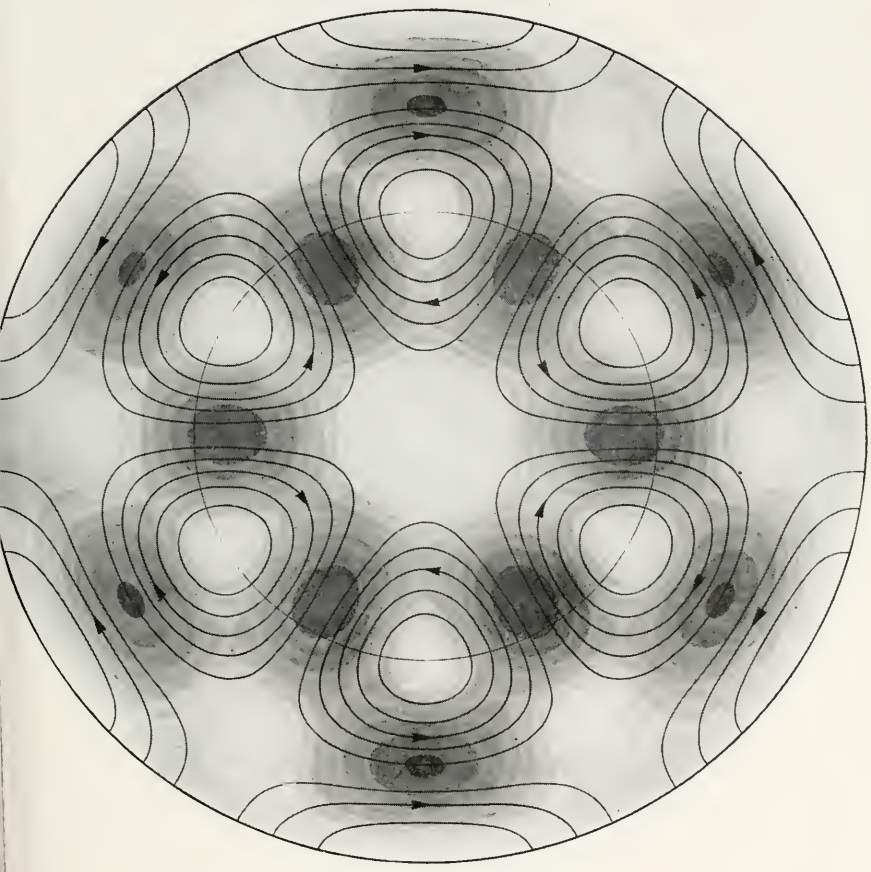


Fig. 46—TE 32 mode.

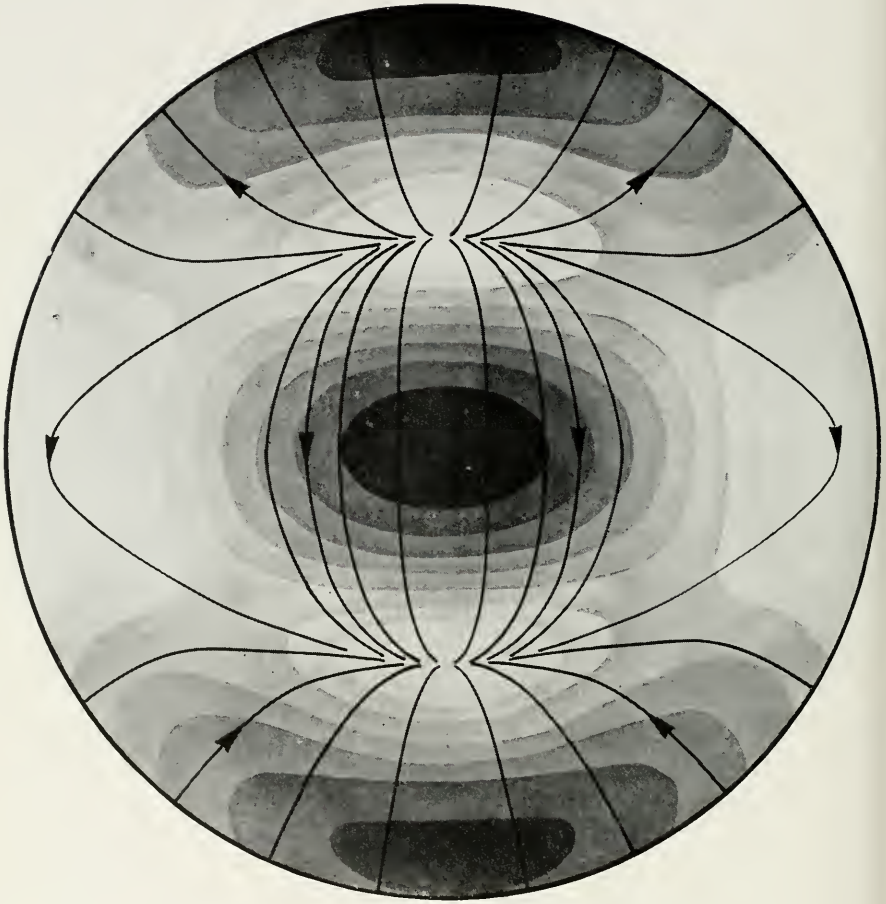


Fig. 47—TM 11 mode.

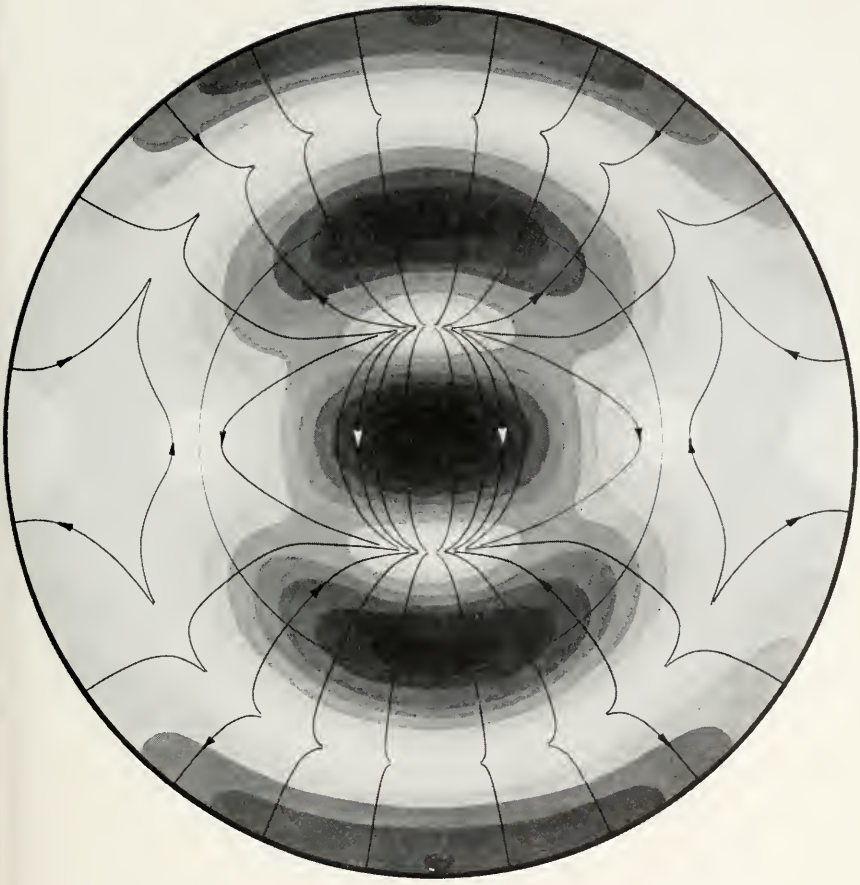


Fig. 48—TM 12 mode.

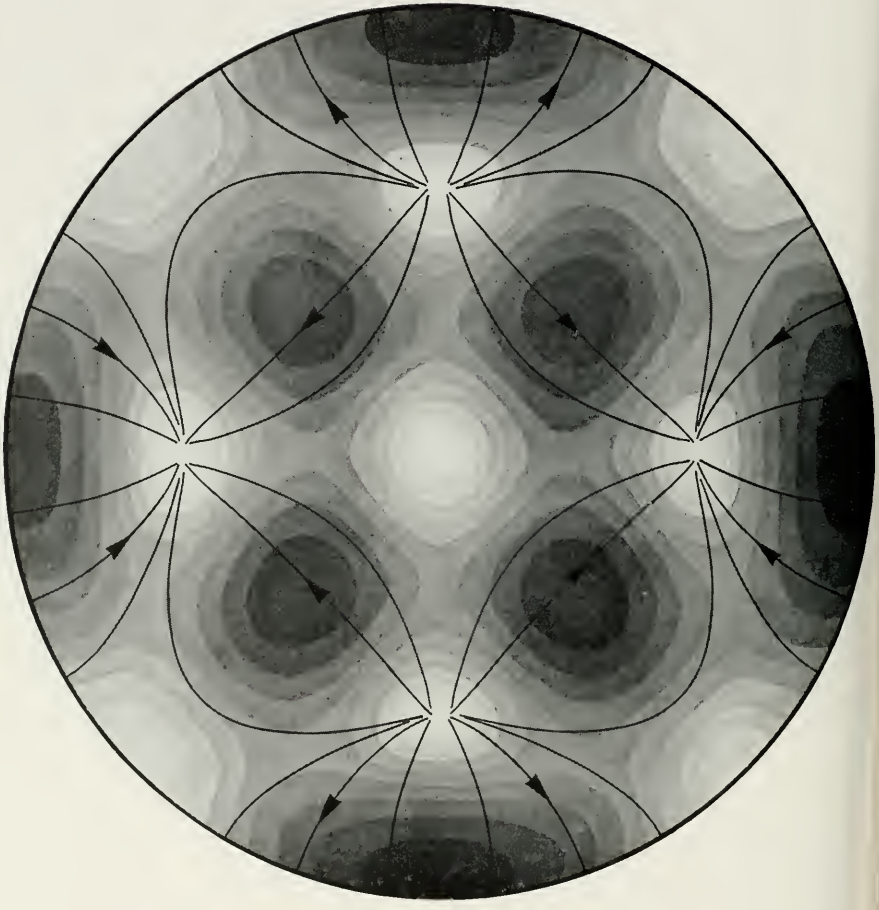


Fig. 49—TM 21 mode

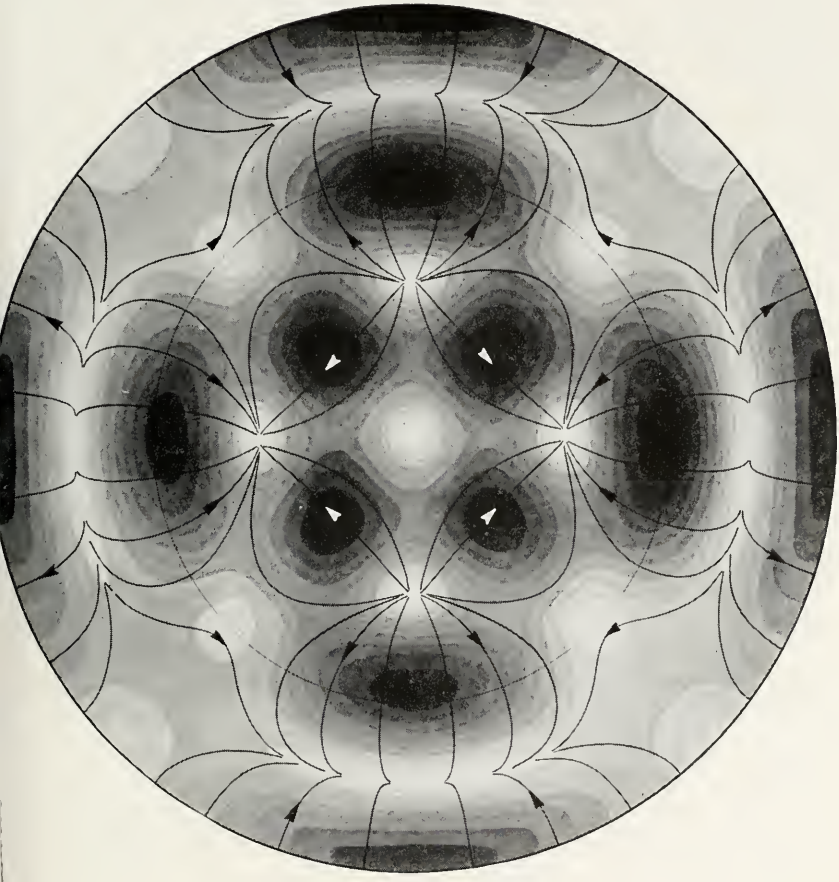


Fig. 50—TM 22 mode.

APPENDIX

INTEGRATION OF $\int_0^x \frac{J_\ell(x)}{J'_\ell(x)}\,dx$

The discussion here is concerned only with integral values of $\ell > 0$. The integral is not simply expressible in terms of known (i.e., tabulated) functions, hence what amounts to a series expansion is used. The method follows Ludinegg[1] who gives the details for $\ell = 1$.

The value of the integrand at $x = 0$ is first discussed. For $\ell = 1$, $J_1(0) = 0$ and $J'_1(0) = 0.5$, hence the integrand has the value zero. For $\ell > 1$, both numerator and denominator are zero, hence the value is indeterminate. Evaluation by $(\ell - 1)$ differentiations of numerator and denominator separately leads to the result that the integrand (and the integral also) is zero at $x = 0$ for all ℓ.

We now introduce a constant p_ℓ and a function $\phi_\ell(x)$ which are such that the following equation is satisfied, at least for a certain range of values of x:

$$J_\ell = -p_\ell\left(J''_\ell - \frac{(\ell - 1)J'_\ell}{x}\right) + \phi_\ell J'_\ell. \tag{1}$$

Denote the desired integral by $F_\ell(x)$, i.e.:

$$F_\ell(x) = \int_0^x \frac{J_\ell(x)}{J'_\ell(x)}\,dx. \tag{2}$$

Then substitution of (1) into (2) yields:

$$F_\ell = -p_\ell\left[\log \frac{J'_\ell}{x^{(\ell-1)}}\right]_0^x + \int_0^x \phi_\ell\,dx. \tag{3}$$

For $x = 0$, $J'_\ell/x^{(\ell-1)}$ is indeterminate, but evaluation by differentiating numerator and denominator separately $(\ell - 1)$ times gives the value $1/2^\ell(\ell - 1)!$

If we can now arrange matters so that ϕ_ℓ remains finite in the range $(0, x)$, its integration can be carried out, a) by expansion into a power series and integration term-by-term, or, b) by numerical integration.

Solving (1) for ϕ_ℓ one obtains

$$\phi_\ell = \frac{J_\ell + p_\ell\left(J''_\ell - \dfrac{(\ell - 1)J'_\ell}{x}\right)}{J'_\ell}. \tag{4}$$

Equation (4) becomes indeterminate at $x = 0$, when $\ell > 1$. Evaluation by differentiating numerator and denominator separately ℓ times shows $\phi_\ell(0) = 0$.

[1] *Hochfrequenztech. u. Elektroak.*, V. 62, pp. 38–44, Aug. 1943.

At the first zero of J'_ℓ (the value of x at a zero of J'_ℓ will be denoted by r), ϕ_ℓ is held finite by choice of the value of p_ℓ. It is clear that (4) becomes indeterminate at $x = r$, if

$$p_\ell = -\frac{J_\ell(r)}{J''_\ell(r)}. \tag{5}$$

Since J_ℓ satisfies the differential equation

$$J''_\ell + \frac{1}{x} J'_\ell + (1 - \ell^2/x^2)J_\ell = 0 \tag{6}$$

and $J'_\ell(r) = 0$, one has by substitution

$$p_\ell = \frac{r^2}{r^2 - \ell^2}. \tag{7}$$

Values of p for several cases are:

ℓ	=	1	2	3	4	1		1
r_1	=	1.841	3.054	4.201	5.318	r_2 = 5.331	r_3 =	8.536
p_ℓ	=	1.418	1.751	2.040	2.303	1.036		1.014
$\phi_\ell(r)$	=	−0.126	−0.286	−0.446	−0.604	−0.180		−0.115

Evaluation of $\phi_\ell(r)$ by the usual process[2] gives:

$$\phi_\ell(r) = \frac{-r\ell(r^2 - \ell^2 - 2\ell)}{(r^2 - \ell^2)^2} \tag{8}$$

Values of $\phi_\ell(r)$ are given in the preceding table.

Since ϕ_ℓ is finite at the origin and at the first zero of J'_ℓ, it may be expanded into a Maclaurin series whose radius of convergence does not, however, exceed the value of x at the second zero of J'_ℓ. Alternatively, by choosing p_ℓ to keep ϕ_ℓ finite at the second (or k^{th}) zero of J'_ℓ it may be expanded into a Taylor series about some point in the interval between the first (or $(k-1)^{th}$) and third (or $(k+1)^{th}$) zeros. Expansions about the origin are given in Table I.

Unfortunately, the convergence of these power series is so slow that they are not very useful. Instead, equation (4) is used to calculate ϕ_ℓ and $\int \phi_\ell \, dx$ is obtained by numerical integration.

With p_ℓ fixed to hold ϕ_ℓ finite at the first root, ζ_1, of J'_ℓ, it is soon found that ϕ_ℓ becomes infinite at the higher roots. This is because different values

[2] Substitute (6) into (4) to eliminate J''_ℓ; differentiate numerator and denominator separately; use (6) to eliminate J_ℓ; allow $x \to r$, using $J'_\ell(r) = 0$ and value of p_ℓ from (7).

of p are required at the different roots, as shown for $\ell = 1$ in the table above. A logical extension would therefore be to make p a function of x such that it takes on the required values at $r_1, r_2, r_3, \cdots$. When this is done and $p\ell(x)$ is introduced into (1) and (2), one has to integrate

$$\int \frac{p(x)J''(x)}{J'(x)}\,dx$$

and this is intractable.[3]

Hence $p(x)$ is made a discontinuous function, such that p has the value p_1 corresponding to r_1 for values of x from zero to a point b_1 between r_1 and r_2; the value p_2 corresponding to r_2 for values of x from b_1 to a point b_2 between r_2 and r_3; and so forth. This introduces discontinuities in ϕ. No discontinuities exist, however, in the function

$$G\ell = e^{-F}\ell \tag{9}$$

which is given in Table II. The calculations were made by Miss F. C. Larkey; numerical integration was according to Weddle's rule.

Within the limits of this tabulation, then, $G\ell$ and $F\ell$ are now considered to be known functions.

TABLE I

POWER SERIES EXPANSIONS OF $\phi_\ell(x)$

$$\phi_1(x) = \left(1 - \frac{3p}{4}\right)x + \left(\frac{1}{4} - \frac{17p}{96}\right)x^3 + \left(\frac{7}{96} - \frac{79p}{1536}\right)x^5 + \cdots$$

$$= -0.063813x - 0.001178x^3 - 0.0000358x^5 - \cdots$$

$$\phi_2(x) = \left(\frac{1}{2} - \frac{p}{3}\right)x + \left(\frac{1}{24} - \frac{7p}{288}\right)x^3 + \left(\frac{5}{1152} - \frac{169p}{17280}\right)x^5 + \cdots$$

$$= +0.15451x + 0.01648x^3 - 0.00580x^5 - \cdots$$

$$\phi_3(x) = \left(\frac{1}{3} - \frac{5p}{24}\right)x + \left(\frac{1}{72} - \frac{41p}{5760}\right)x^3 + \left(\frac{13}{17280} - \frac{103p}{276480}\right)x^5 + \cdots$$

$$= +0.12210x + 0.00667x^3 + 0.00375x^5 - \cdots .$$

[3] Unless $p = b + cJ'$ (b and c constants), which is not of any use.

TABLE II

VALUES OF $F_1(x) = \int_0^x \frac{J_1(x)}{J_1'(x)}\,dx$; $G_1(x) = e^{-F_1}$

$F_1(x)$

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0	0050	0201	0455	0816	1291	1887	2616	3493	4539
1	5782	7261	9036	1.1192	1.3874	1.7336	2.2103	2.9577	4.6961	4.1846
2	2.7727	2.0801	1.6199	1.2775	1.0073	7864	6018	4454	3117	1970
3	0987	0147	−0564	−1157	−1640	−2018	−2296	−2475	−2556	−2537
4	−2416	−2188	−1845	−1377	−0769	0	+0960	2153	3646	5549
5	8060	1.1595	1.7307	3.2014	2.3851	1.4478	9635	6373	3939	2024
6	0470	−0812	−1879	−2768	−3506	−4111	−4594	−4966	−5233	−5398
7	−5463	−5429	−5292	−5049	−4693	−4214	−3598	−2826	−1868	−0685
8	+0789	2657	5107	8530	1.3992	2.7313	2.1565	1.1974	7154	3942
9	1562	−0300	−1802	−3034	−4053	−4897	−5590	−6150	−6591	−6921

$G_1(x)$

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	1.0000	9950	9801	9555	9216	8789	8280	7698	7052	6351
1	5609	4838	4051	3265	2497	1766	1097	0519	0091	0152
2	0625	1249	1979	2787	3652	4555	5478	6406	7322	8212
3	9060	9854	1.0580	1.1226	1.1781	1.2236	1.2581	1.2808	1.2912	1.2888
4	1.2733	1.2445	1.2026	1.1476	1.0799	1.0000	9085	8063	6945	5741
5	4467	3136	1772	0407	0921	2351	3816	5287	6744	8168
6	9541	1.0846	1.2067	1.3190	1.4200	1.5084	1.5831	1.6432	1.6877	1.7157
7	1.7269	1.7209	1.6976	1.6568	1.5989	1.5241	1.4331	1.3265	1.2054	1.0709
8	9241	7667	6091	4261	2468	0613	1157	3020	4890	6742
9	8554	1.0304	1.1974	1.3545	1.4998	1.6318	1.7489	1.8497	1.9330	1.9978

VALUES OF $F_2(x) = \int_0^x \frac{J_2(x)}{J_2(x)}\,dx$; $G_2(x) = e^{-F_2}$

$F_2(x)$

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0	0025	0100	0226	0403	0632	0914	1251	1645	2097
1	2612	3192	3840	4563	5365	6253	7236	8323	9528	1.0866
2	1.2357	1.4008	1.5913	1.8061	2.0541	2.3456	2.6972	3.1380	3.7263	4.6110
3	6.4527	6.7644	4.7528	3.8572	3.2808	2.8597	2.5316	2.2658	2.0451	1.8590
4	1.7002	1.5641	1.4470	1.3466	1.2607	1.1881	1.1275	1.0783	1.0396	1.0112
5	9928	9843	9858	9974	1.0196	1.0530	1.0985	1.1573	1.2311	1.3223
6	1.4345	1.5726	1.7447	1.9640	2.2555	2.6743	3.3910	6.5119	3.5122	2.7144
7	2.2595	1.9432	1.7034	1.5131	1.3579	1.2294	1.1223	1.0328	.9586	.8977
8	.8490	.8115	.7846	.7679	.7612	.7645	.7779	.8020	.8372	.8845
9	.9452	1.0212	1.1149	1.2301	1.3725	1.5512	1.7817	2.0950	2.5660	3.4864

$G_2(x)$

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	1.0000	9975	9900	9777	9605	9388	9127	8824	8483	8108
1	7701	7267	6811	6336	5848	5351	4850	4350	3856	3373
2	2906	2459	2036	1643	1282	0958	0674	0434	0241	0099
3	0017	0012	0086	0211	0376	0573	0795	1037	1294	1558
4	1826	2093	2353	2601	2834	3048	3238	3402	3536	3638
5	3705	3737	3731	3688	3607	3489	3334	3143	2920	2665
6	2383	2075	1747	1403	1048	0690	0337	0015	0298	0662
7	1044	1432	1821	2202	2572	2925	3255	3560	3834	4075
8	4278	4442	4563	4640	4671	4656	4593	4484	4329	4129
9	8886	3602	3280	2923	2535	2120	1683	1231	0768	0306

VALUES OF $F_3(x) = \int_0^x \frac{J_3(x)}{J_3'(x)}\,dx$; $G_3(x) = e^{-F_3}$

$F_3(x)$

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0	0017	0067	0152	0268	0420	0604	0826	1081	1373
1	1703	2070	2476	2922	3410	3942	4518	5141	5814	6539
2	7319	8158	9060	1.0028	1.1070	1.2192	1.3401	1.4706	1.6118	1.7650
3	1.9321	2.1150	2.3165	2.5402	2.7908	3.0752	3.4034	3.7905	4.2624	4.8669
4	5.7117	7.1373	16.2303	7.2383	5.8409	5.0409	4.4852	4.0643	3.7292	3.4543
5	3.2239	3.0282	2.8605	2.7160	2.5913	2.4838	2.3914	2.3128	2.2467	2.1922
6	2.1487	2.1156	2.0927	2.0798	2.0768	2.0838	2.1012	2.1293	2.1685	2.2208
7	2.2864	2.3674	2.4664	2.5868	2.7340	2.9159	3.1460	3.4491	3.8790	4.5950
8	6.9408	4.9414	4.0348	3.5348	3.1912	2.9324	2.7276	2.5608	2.4227	2.3074
9	2.2108	2.1302	2.0637	2.0097	1.9676	1.9361	1.9147	1.9036	1.9025	1.9115

$G_3(x)$

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	1.0000	9983	9933	9849	9734	9589	9413	9208	8975	8717
1	8434	8130	7806	7466	7110	6742	6365	5980	5591	5200
2	4810	4423	4041	3668	3305	2955	2618	2298	1995	1712
3	1448	1206	0986	0789	0614	0462	0333	0226	0141	0077
4	0033	0008	0000	0007	0029	0065	0113	0172	0240	0316
5	0398	0484	0572	0661	0749	0834	0915	0990	1057	1117
6	1166	1206	1233	1250	1253	1244	1223	1189	1143	1085
7	1016	0937	0849	0753	0650	0542	0430	0318	0207	0101
8	0010	0071	0177	0292	0411	0533	0654	0772	0887	0995
9	1096	1188	1270	1340	1398	1443	1474	1490	1492	1479

Table III

Bessel Functions of The First Kind

$J_0(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+1.0	9975	9900	9776	9604	+9385	9120	8812	8463	8075
1	+7652	7196	6711	6201	5669	+5118	4554	3980	3400	2818
2	+2239	1666	1104	0555	0025	−0484	0968	1424	1850	2243
3	−2601	2921	3202	3443	3643	−3801	3918	3992	4026	4018
4	−3971	3887	3766	3610	3423	−3205	2961	2693	2404	2097
5	−1776	1443	1103	0758	0412	−0068	+0270	+0599	+0917	+1220
6	+1506	1773	2017	2238	2433	+2601	2740	2851	2931	2981
7	+3001	2991	2951	2882	2786	+2663	2516	2346	2154	1944
8	+1717	1475	1222	0960	0692	+0419	0146	−0125	−0392	−0653
9	−0903	1142	1367	1577	1768	−1939	2090	2218	2323	2403

$J_1(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0499	0995	1483	1960	+2423	2867	3290	3688	4059
1	+4401	4709	4983	5220	5419	+5579	5699	5778	5815	5812
2	+5767	5683	5560	5399	5202	+4971	4708	4416	4097	3754
3	+3391	3009	2613	2207	1792	+1374	0955	0538	0128	−0272
4	−0660	1033	1386	1719	2028	−2311	2566	2791	2985	3147
5	−3276	3371	3432	3460	3453	−3414	3343	3241	3110	2951
6	−2767	2559	2329	2081	1816	−1538	1250	0953	0652	0349
7	−0047	+0252	+0543	+0826	+1096	+1352	1592	1813	2014	2192
8	+2346	2476	2580	2657	2708	+2731	2728	2697	2641	2559
9	+2453	2324	2174	2004	1816	+1613	1395	1166	0928	0684

$J_2(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0012	0050	0112	0197	0306	0437	0588	0758	0946
1	+1149	1366	1593	1830	2074	2321	2570	2817	3061	3299
2	+3528	3746	3951	4139	4310	4461	4590	4696	4777	4832
3	+4861	4862	4835	4780	4697	4586	4448	4283	4093	3879
4	+3641	3383	3105	2811	2501	2178	1846	1506	1161	0813
5	+0466	0121	−0217	−0547	−0867	−1173	1464	1737	1990	2221
6	−2429	2612	2769	2899	3001	3074	3119	3135	3123	3082
7	−3014	2920	2800	2656	2490	2303	2097	1875	1638	1389
8	−1130	0864	0593	0320	0047	+0223	0488	0745	0993	1228
9	+1448	1653	1840	2008	2154	2279	2380	2458	2512	2542

$J_3(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0	0002	0006	0013	0026	0044	0069	0102	0144
1	+0196	0257	0329	0411	0505	0610	0725	0851	0988	1134
2	+1289	1453	1623	1800	1981	2166	2353	2540	2727	2911
3	+3091	3264	3431	3588	3734	3868	3988	4092	4180	4250
4	+4302	4333	4344	4333	4301	4247	4171	4072	3952	3811
5	+3648	3466	3265	3046	2811	2561	2298	2023	1738	1446
6	+1148	0846	0543	0240	−0059	−0353	0641	0918	1185	1438
7	−1676	1896	2099	2281	2442	2581	2696	2787	2853	2895
8	−2911	2903	2869	2811	2730	2626	2501	2355	2190	2007
9	−1809	1598	1374	1141	0900	0653	0403	0153	+0097	+0343

$J_4(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0	0	0	0001	0002	0003	0006	0010	0016
1	+0025	0036	0050	0068	0091	0118	0150	0188	0232	0283
2	+0340	0405	0476	0556	0643	0738	0840	0950	1067	1190
3	+1320	1456	1597	1743	1892	2044	2198	2353	2507	2661
4	+2811	2958	3100	3236	3365	3484	3594	3693	3780	3853
5	+3912	3956	3985	3996	3991	3967	3926	3866	3788	3691
6	+3576	3444	3294	3128	2945	2748	2537	2313	2077	1832
7	+1578	1317	1051	0781	0510	0238	−0031	−0297	−0557	−0810
8	1054	1286	1507	1713	1903	2077	2233	2369	2485	2581
9	−2655	2707	2736	2743	2728	2691	2633	2553	2453	2334

$J_5(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0	0	0	0	0	0	0	0001	0001
1	+0002	0004	0006	0009	0013	0018	0025	0033	0043	0055
2	+0070	0088	0109	0134	0162	0195	0232	0274	0321	0373
3	+0430	0493	0562	0637	0718	0804	0897	0995	1098	1207
4	+1321	1439	1561	1687	1816	1947	2080	2214	2347	2480
5	+2611	2740	2865	2986	3101	3209	3310	3403	3486	3559
6	+3621	3671	3708	3731	3741	3736	3716	3680	3629	3562
7	+3479	3380	3266	3137	2993	2835	2663	2478	2282	2075
8	+1858	1632	1399	1161	0918	0671	0424	0176	−0070	−0313
9	−0550	0782	1005	1219	1422	1613	1790	1953	2099	2229

$J_6(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0001	0001	0002	0002	0003	0005	0007	0009
2	0012	0016	0021	0027	0034	0042	0052	0065	0079	0095
3	0114	0136	0160	0188	0219	0254	0293	0336	0383	0435
4	0491	0552	0617	0688	0763	0843	0927	1017	1111	1209
5	1310	1416	1525	1637	1751	1868	1986	2104	2223	2341
6	2458	2574	2686	2795	2900	2999	3093	3180	3259	3330
7	3392	3444	3486	3516	3535	3541	3535	3516	3483	3436
8	3376	3301	3213	3111	2996	2867	2725	2571	2406	2230
9	2043	1847	1644	1432	1215	0993	0768	0540	0311	0082

$J_7(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0001	0001	0001
2	0002	0002	0003	0004	0006	0008	0010	0013	0016	0020
3	0025	0031	0038	0047	0056	0067	0080	0095	0112	0130
4	0152	0176	0202	0232	0264	0300	0340	0382	0429	0479
5	0534	0592	0654	0721	0791	0866	0945	1027	1113	1203
6	1296	1392	1491	1592	1696	1801	1908	2015	2122	2230
7	2336	2441	2543	2643	2739	2832	2919	3001	3076	3145
8	3206	3259	3303	3337	3362	3376	3379	3371	3351	3319
9	3275	3218	3149	3068	2974	2868	2750	2620	2480	2328

$J_1'(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+5000	4981	4925	4832	4703	4539	4342	4112	3852	3565
1	+3251	2915	2559	2185	1798	1399	0992	0581	0169	−0241
2	−0645	1040	1423	1792	2142	2472	2779	3060	3314	3538
3	−3731	3891	4019	4112	4170	4194	4183	4138	4059	3948
4	−3806	3635	3435	3210	2962	2692	2404	2100	1782	1455
5	−1121	0782	0443	0105	+0227	+0552	0867	1168	1453	1721
6	+1968	2192	2393	2568	2717	2838	2930	2993	3027	3032
7	+3007	2955	2875	2769	2638	2483	2307	2110	1896	1666
8	+1423	1169	0908	0640	0369	0098	−0171	−0435	−0692	−0940
9	−1176	1398	1604	1792	1961	2109	2235	2338	2417	2472

$J_2'(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0250	0497	0739	0974	1199	1412	1610	1793	1958
1	+2102	2226	2327	2404	2457	2485	2487	2463	2414	2339
2	+2239	2115	1968	1799	1610	1402	1178	0938	0685	0422
3	+0150	−0128	−0409	−0691	−0971	−1247	1516	1777	2026	2261
4	−2481	2683	2865	3026	3165	3279	3368	3432	3469	3479
5	−3462	3419	3349	3253	3132	2988	2821	2632	2424	2199
6	−1957	1702	1436	1161	0879	0592	0305	0018	+0266	+0544
7	+0814	1074	1321	1553	1769	1967	2144	2300	2434	2543
8	+2629	2689	2725	2734	2719	2679	2614	2526	2415	2283
9	+2131	1961	1774	1572	1358	1133	0899	0659	0416	0170

$J_3'(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0006	0025	0056	0098	0152	0217	0291	0374	0465
1	+0562	0665	0772	0881	0991	1102	1210	1315	1415	1508
2	+1594	1671	1737	1792	1833	1861	1875	1873	1855	1821
3	+1770	1703	1619	1519	1403	1271	1125	0965	0793	0609
4	+0415	0212	0003	−0213	−0432	−0653	0874	1094	1310	1520
5	−1723	1918	2101	2272	2429	2570	2695	2801	2889	2956
6	−3003	3028	3031	3013	2973	2911	2828	2724	2600	2457
7	−2296	2118	1925	1719	1500	1270	1033	0789	0540	0289
8	−0038	+0211	+0457	+0696	+0928	+1150	1360	1557	1739	1904
9	+2052	2180	2288	2376	2441	2485	2507	2506	2483	2438

$J_4'(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0	0001	0003	0007	0013	0022	0034	0051	0071
1	+0097	0126	0161	0201	0246	0296	0350	0409	0473	0539
2	+0610	0682	0757	0833	0909	0985	1060	1133	1203	1269
3	+1330	1385	1434	1475	1508	1532	1545	1549	1541	1522
4	+1490	1447	1391	1323	1243	1150	1045	0929	0802	0665
5	+0518	0363	0200	0030	−0145	−0324	0506	0690	0874	1057
6	−1237	1412	1582	1745	1900	2045	2178	2299	2407	2500
7	−2577	2638	2683	2709	2718	2708	2679	2633	2568	2485
8	−2385	2267	2134	1986	1824	1649	1462	1265	1060	0847
9	−0629	0408	0184	+0039	+0261	+0480	0694	0900	1098	1286

$J_5'(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0	0	0	0	0001	0002	0003	0005	0008
1	+0012	0018	0025	0034	0045	0058	0073	0092	0113	0137
2	+0164	0194	0228	0265	0305	0348	0394	0443	0494	0548
3	+0603	0660	0718	0777	0836	0895	0952	1008	1062	1113
4	+1160	1203	1242	1274	1301	1321	1333	1338	1335	1322
5	+1301	1270	1230	1180	1120	1050	0970	0881	0782	0675
6	+0559	0435	0304	0166	0023	−0126	0278	0433	0591	0749
7	−0907	1064	1217	1368	1513	1652	1783	1906	2020	2123
8	−2215	2294	2360	2412	2449	2472	2479	2470	2446	2405
9	−2349	2277	2190	2088	1972	1842	1700	1546	1382	1208

$J_6'(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	+0	0	0	0	0	0	0	0	0	0001
1	+0001	0002	0003	0004	0006	0009	0012	0016	0021	0027
2	+0034	0043	0053	0065	0078	0094	0111	0130	0152	0176
3	+0202	0231	0262	0295	0331	0368	0408	0450	0493	0538
4	+0585	0632	0680	0728	0776	0823	0870	0916	0959	1000
5	+1039	1074	1105	1132	1155	1172	1183	1188	1187	1178
6	+1163	1139	1108	1069	1022	0967	0904	0833	0753	0666
7	+0572	0470	0362	0247	0127	0002	−0128	−0261	−0397	−0535
8	−0674	0813	0952	1088	1222	1352	1478	1597	1710	1816
9	−1912	2000	2077	2143	2198	2240	2270	2287	2290	2279

Table IV

Relative Radius for Maximum of ρI_μ^2

Mode			
TE 11	.737		
12	.982	.254	
13	.993	.613	.159
21	.894		
22	.988	.407	
23	.995	.664	.274
31	.937		
32	.991	.491	
41	.956		
42	.993	.548	
51	.967		
61	.974		
TM 01	.901		
02	.983	.393	
03	.993	.627	.250
11	.961		
12	.989	.525	
13	.995	.682	.362
21	.977		
22	.992	.596	
31	.984		
32	.994	.643	
41	.988		
51	.990		
61	.992		

First and Second Order Equations for Piezoelectric Crystals Expressed in Tensor Form

By W. P. MASON

Introduction

ANEOLOTROPIC substances have been used for a wide variety of elastic piezoelectric, dielectric, pyroelectric, temperature expansive, piezo-optic and electro-optic effects. While most of these effects may be found treated in various publications[1] there does not appear to be any integrated treatment of them by the tensor method which greatly simplifies the method of writing and manipulating the relations between fundamental quantities. Other short hand methods such as the matrix method[2] can also be used for all the linear effects, but for second order effects involving tensors higher than rank four, tensor methods are essential. Accordingly, it is the purpose of this paper to present such a derivation. The notation used is that agreed upon by a committee of piezoelectric experts under the auspices of the Institute of Radio Engineers.

In the first part the definition of stress and strain are given and their interrelation, the generalized Hookes law is discussed. The modifications caused by adiabatic conditions are considered. When electric fields, stresses, and temperature changes are applied, there are nine first order effects each of which requires a tensor to express the resulting constants. The effects are the elastic effect, the direct and inverse piezoelectric effects, the temperature expansion effect, the dielectric effect, the pyroelectric effect, the heat of deformation, the electrocaloric effect, and the specific heat. There are three relations between these nine effects. Making use of the tensor transformation of axes, the results of the symmetries existing for the 32 types of crystals are investigated and the possible constants are derived for these nine effects.

Methods are discussed for measuring these properties for all 32 crystal classes. By measuring the constants of a specified number of oriented cuts for each crystal class, vibrating in longitudinal and shear modes, all of the elastic, dielectric and piezoelectric constants can be obtained. Methods for calculating the properties of the oriented cuts are given and for deriving the fundamental constants from these measurements.

[1] For example Voigt, "Lehrbuch der Kristall Physik," B. Teubner, 1910; Wooster, "Crystal Physics," Cambridge Press, 1938; Cady "Piezo-electricity" McGraw Hill, 1946.

[2] The matrix method is well described by W. L. Bond "The Mathematics of the Physical Properties of Crystals," B. S. T. J., Vol. 22, pp. 1–72, 1943.

Second order effects are also considered. These effects (neglecting second order temperature effects) are elastic constants whose values depend on the applied stress and the electric displacement, the electrostrictive effect, piezoelectric constants that depend on the applied stress, the piezo-optical effect and the electro-optical effect. These second order equations can also be used to discuss the changes that occur in ferroelectric type crystals such as Rochelle Salt, for which between the temperature of $-18°$C. and $+24°$C., a spontaneous polarization occurs along one direction in the crystal. This spontaneous polarization gives rise to a first order piezoelectric deformation and to second order electrostrictive effects. It produces changes in the elastic constants, the piezoelectric constants and the dielectric constants. Some measurements have been made for Rochelle Salt evaluating these second order constants.

Mueller in his theory of Rochelle Salt considers that the crystal changes from an orthorhombic crystal to a monoclinic crystal when it becomes spontaneously polarized. An alternate view developed here is that all of the new constants created by the spontaneous polarization are the result of second order effects in the orthorhombic crystal. As shown in section 7 these produce new constants proportional to the square of the spontaneous polarization which are the ones existing in a monoclinic crystal. On this view "morphic" effects are second order effects produced by the spontaneous polarization.

1. Stress and Strain Relations in Aeolotropic Crystals

I.I. *Specification of Stress*

The stresses exerted on any elementary cube of material with its edges along the three rectangular axes X, Y and Z can be specified by considering the stresses on each face of the cube illustrated by Fig. 1. The total stress acting on the face ABCD normal to the X axis can be represented by a resultant force R, with its center of application at the center of the face, plus a couple which takes account of the variation of the stress across the face. The force R is directed outward, since a stress is considered positive if it exerts a tension. As the face is shrunk in size, the force R will be proportional to the area of the face, while the couple will vary as the cube of the dimension. Hence in the limit the couple can be neglected with respect to the force R. The stress (force per unit area) due to R can be resolved into three components along the three axes to which we give the designation

$$T_{xx_2}, \qquad T_{yx_2}, \qquad T_{zx_2}. \tag{1}$$

Here the first letter designates the direction of the stress component and the second letter x_2 denotes the second face of the cube normal to the X axis. Similarly for the first X face OEFG, the stress resultant can be resolved

into the components T_{xx_1}, T_{yx_1}, T_{zx_1}, which are oppositely directed to those of the second face. The remaining stress components on the other four faces have the designation

$$\begin{array}{llll} \text{Face OABE} & T_{xy_1}, & T_{yy_1}, & T_{zy_1} \\ \text{CFGD} & T_{xy_2}, & T_{yy_2}, & T_{zy_2} \\ \text{OADG} & T_{xz_1}, & T_{yz_1}, & T_{zz_1} \\ \text{BCFE} & T_{xz_2}, & T_{yz_2}, & T_{zz_2}. \end{array} \tag{2}$$

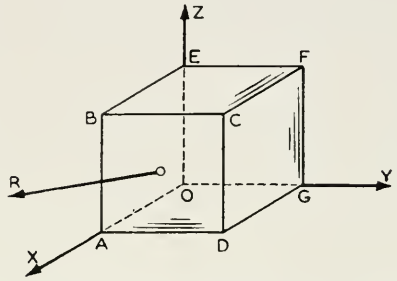


Fig. 1.—Cube showing method for specifying stresses.

The resultant force in the X direction is obtained by summing all the forces with components in the X direction or

$$F_x = (T_{xx_1} - T_{xx_2})\, dydz + (T_{xy_1} - T_{xy_2})\, dxdz + (T_{xz_1} - T_{xz_2})\, dxdy. \tag{3}$$

But

$$\begin{aligned} T_{xx_2} &= -T_{xx_1} + \frac{\partial T_{xx}}{\partial x}\, dx; \qquad T_{xy_2} = -T_{xy_1} \\ &\quad + \frac{\partial T_{xy}}{\partial y}\, dy; \qquad T_{xz_2} = -T_{xz_1} + \frac{\partial T_{xz}}{\partial z}\, dz \end{aligned} \tag{4}$$

and equation (3) can be written in the form

$$F_x = -\left(\frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{xy}}{\partial y} + \frac{\partial T_{xz}}{\partial z}\right) dx\, dy\, dz. \tag{5}$$

Similarly the resultant forces in the other directions are

$$\begin{aligned} F_y &= -\left(\frac{\partial T_{yx}}{\partial x} + \frac{\partial T_{yy}}{\partial y} + \frac{\partial T_{yz}}{\partial z}\right) dx\, dy\, dz \\ F_z &= -\left(\frac{\partial T_{zx}}{\partial x} + \frac{\partial T_{zy}}{\partial y} + \frac{\partial T_{zz}}{\partial z}\right) dx\, dy\, dz. \end{aligned} \tag{6}$$

We call the components

$$\begin{vmatrix} T_{xx}, & T_{xy}, & T_{xz} \\ T_{yx}, & T_{yy}, & T_{yz} \\ T_{zx}, & T_{zy}, & T_{zz} \end{vmatrix} = \begin{vmatrix} T_{11}, & T_{12}, & T_{13} \\ T_{21}, & T_{22}, & T_{23} \\ T_{31}, & T_{32}, & T_{33} \end{vmatrix} \tag{7}$$

the stress components exerted on the elementary cube which tend to deform it. The rate of change of these stresses determines the resultant force on the cube. The second form of (7) is commonly used when the stresses are considered as a second rank tensor.

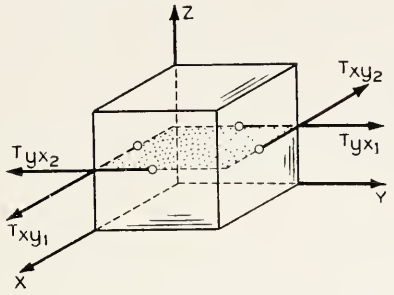


Fig. 2.—Shearing stresses exerted on a cube.

It can be shown that there is a relation between 3 pairs of these components, namely

$$T_{xy} = T_{yx}\,; \qquad T_{xz} = T_{zx}\,; \qquad T_{yz} = T_{zy}\,. \tag{8}$$

To show this consider Fig. 2 which shows the stresses tending to rotate the elementary cube about the Z axis. The stresses T_{yx_2} and T_{yx_1} tend to rotate the cube about the Z axis by producing the couple

$$\frac{T_{yx}\,dx\,dy\,dz}{2}\,. \tag{9}$$

The stresses T_{xy_1} and T_{xy_2} produce a couple tending to cause a rotation in the opposite direction so that

$$\tfrac{1}{2}\,(T_{yx} - T_{xy})\,dx\,dy\,dz = \text{couple} = I\ddot{\omega}_z \tag{10}$$

is the total couple tending to produce a rotation around the Z axis. But from dynamics, it is known that this couple is equal to the product of the moment of inertia of the section times the angular acceleration. This moment of inertia of the section is proportional to the fourth power of the cube edge and the angular acceleration is finite. Hence as the cube edge

approaches zero, the right hand side of (10) is one order smaller than the left hand side and hence

$$T_{yx} = T_{xy}. \tag{11}$$

The same argument applies to the other terms. Hence the stress components of (7) can be written in the symmetrical form

$$\begin{vmatrix} T_{xx}, & T_{xy}, & T_{xz} \\ T_{xy}, & T_{yy}, & T_{yz} \\ T_{xz}, & T_{yz}, & T_{zz} \end{vmatrix} = \begin{vmatrix} T_{11}, & T_{12}, & T_{13} \\ T_{12}, & T_{22}, & T_{23} \\ T_{13}, & T_{23}, & T_{33} \end{vmatrix} = \begin{vmatrix} T_1, & T_6, & T_5 \\ T_6, & T_2, & T_4 \\ T_5, & T_4, & T_3 \end{vmatrix} \tag{12}$$

The last form is a short hand method for reducing the number of indices in the stress tensor. The reduced indices 1 to 6, correspond to the tensor indices if we replace

11 by 1; 22 by 2; 33 by 3; 23 by 4; 13 by 5; 12 by 6.

This last method is the most common way for writing the stresses.

1.2 *Strain Components*

The types of strain present in a body can be specified by considering two points P and Q of a medium, and calculating their separation in the strained condition. Let us consider the point P at the origin of coordinates and the point Q having the coordinates x, y and z as shown by Fig. 3. Upon strain-

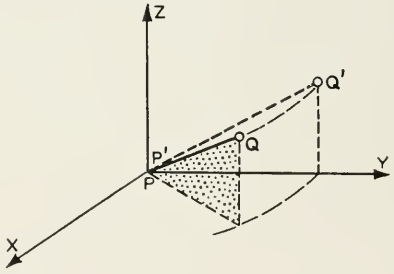


Fig. 3.—Change in length and position of a line due to strain in a solid body.

ing the body, the points change to the positions P', Q'. In order to specify the strains, we have to calculate the difference in length after straining, or have to evaluate the distance $P'Q' - P\,Q$. After the material has stretched the point P' will have the coordinates ξ_1, η_1, ζ_1, while Q' will have the coordinates $x + \xi_2$; $y + \eta_2$; $z + \zeta_2$. But the displacement is a continuous function of the coordinates x, y and z so that we have

$$\xi_2 = \xi_1 + \frac{\partial \xi}{\partial x} x + \frac{\partial \xi}{\partial y} y + \frac{\partial \xi}{\partial z} z.$$

Similarly

$$\eta_2 = \eta_1 + \frac{\partial \eta}{\partial x} x + \frac{\partial \eta}{\partial y} y + \frac{\partial \eta}{\partial z} z$$
$$\zeta_2 = \zeta_1 + \frac{\partial \zeta}{\partial x} x + \frac{\partial \zeta}{\partial y} y + \frac{\partial \zeta}{\partial z} z. \tag{13}$$

Hence subtracting the two lengths, we find that the increases in separation in the three directions are

$$\delta_x = x \frac{\partial \xi}{\partial x} + y \frac{\partial \xi}{\partial y} + z \frac{\partial \xi}{\partial z}$$
$$\delta_y = x \frac{\partial \eta}{\partial x} + y \frac{\partial \eta}{\partial y} + z \frac{\partial \eta}{\partial z} \tag{14}$$
$$\delta_z = x \frac{\partial \zeta}{\partial x} + y \frac{\partial \zeta}{\partial y} + z \frac{\partial \zeta}{\partial z}.$$

The net elongation of the line in the x direction is $x \frac{\partial \xi}{\partial x}$ and the elongation per unit length is $\frac{\partial \xi}{\partial x}$ which is defined as the linear strain in the x direction. We have therefore that the linear strains in the x, y and z directions are

$$S_1 = \frac{\partial \xi}{\partial x}; \qquad S_2 = \frac{\partial \eta}{\partial y}; \qquad S_3 = \frac{\partial \zeta}{\partial z}. \tag{15}$$

The remaining strain coefficients are usually defined as

$$S_4 = \frac{\partial \zeta}{\partial y} + \frac{\partial \eta}{\partial z}; \qquad S_5 = \frac{\partial \xi}{\partial z} + \frac{\partial \zeta}{\partial x}; \qquad S_6 = \frac{\partial \eta}{\partial x} + \frac{\partial \xi}{\partial y} \tag{16}$$

and the rotation coefficients by the equations

$$\omega_x = \frac{\partial \zeta}{\partial y} - \frac{\partial \eta}{\partial z}; \qquad \omega_y = \frac{\partial \xi}{\partial z} - \frac{\partial \zeta}{\partial x}; \qquad \omega_z = \frac{\partial \eta}{\partial x} - \frac{\partial \xi}{\partial y}. \tag{17}$$

Hence the relative displacement of any two points can be expressed as

$$\delta_x = xS_1 + y\left(\frac{S_6 - \omega_z}{2}\right) + z\left(\frac{S_5 + \omega_y}{2}\right)$$
$$\delta_y = x\left(\frac{S_6 + \omega_z}{2}\right) + yS_2 + z\left(\frac{S_4 - \omega_x}{2}\right) \tag{18}$$
$$\delta_z = x\left(\frac{S_5 - \omega_y}{2}\right) + y\left(\frac{S_4 + \omega_x}{2}\right) + zS_3$$

which represents the most general type of displacement that the line $P\,Q$ can undergo.

As discussed in section 4 the definition of the shearing strains given by equation (16) does not allow them to be represented as part of a tensor. If however we defined the shearing strains as

$$2S_{23} = S_4 = \left(\frac{\partial \zeta}{\partial y} + \frac{\partial \eta}{\partial z}\right); \qquad 2S_{13} = S_5 = \frac{\partial \xi}{\partial z} + \frac{\partial \zeta}{\partial x}; \qquad 2S_{12} = S_6 = \frac{\partial \eta}{\partial x} + \frac{\partial \xi}{\partial y} \tag{19}$$

they can be expressed in the form of a symmetrical tensor

$$\begin{vmatrix} S_{11} & S_{12} & S_{13} \\ S_{12} & S_{22} & S_{23} \\ S_{13} & S_{23} & S_{33} \end{vmatrix} = \begin{vmatrix} S_1 & \frac{S_6}{2} & \frac{S_5}{2} \\ \frac{S_6}{2} & S_2 & \frac{S_4}{2} \\ \frac{S_5}{2} & \frac{S_4}{2} & S_3 \end{vmatrix} \tag{20}$$

For an element suffering a shearing strain $S_6 = 2S_{12}$ only, the displacement along x is proportional to y, while the displacement along y is proportional to the x dimension. A cubic element of volume will be strained into a rhombic form, as shown by Fig. 4, and the cosine of the resulting angle θ

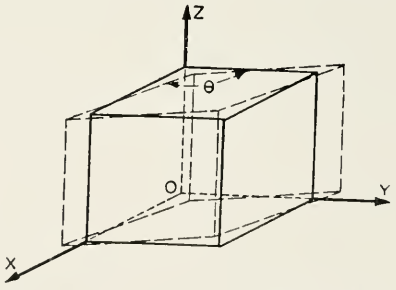


Fig. 4.—Distortion due to a shearing strain.

measures the shearing deformation. For an element suffering a rotation ω_z only, the displacement along x is proportional to y and in the negative y direction, while the displacement along y is in the positive x direction. Hence a rectangle has the displacement shown by Fig. 5, which is a pure rotation of the body without change of form, about the z axis. For any

body in equilibrium or in nonrotational vibration, the ω's can be set equal to zero.

The total potential energy stored in a general distortion can be calculated as the sum of the energies due to the distortion of the various modes. For example in expanding the cube in the x direction by an amount $\frac{\partial \dot{\xi}}{\partial x} dx = S_1 dx$, the work done is the force times the displacement. The force wil

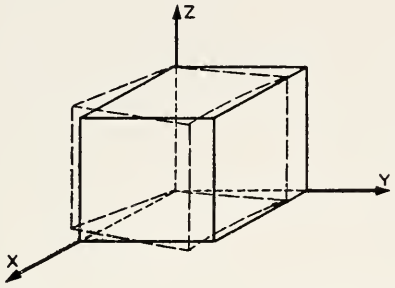


Fig. 5.—A rotation of a solid body.

be the force T_1 and will be $T_1 dy dz$. Hence the potential energy stored in this distortion is

$$T_1 dS_1 dx dy dz$$

For a shearing stress T_6 of the type shown by Fig. 4 the displacement $\frac{dS_6 dx}{2}$ times the force $T_6 dy dz$ and the displacement $\frac{dS_6 dy}{2}$ times the force $T_6 dx dz$ equals the stored energy or

$$\Delta PE_6 = \tfrac{1}{2} (dS_6 T_6 + dS_6 T_6) dx dy dz = dS_6 T_6 dx dy dz.$$

Hence for all modes of motion the stored potential energy is equal to

$$\Delta PE = [T_1 dS_1 + T_2 dS_2 + T_3 dS_3 + T_4 dS_4 + T_5 dS_5 + T_6 dS_6] dx dy dz. \quad (21)$$

1.3 *Generalized Hooke's Law*

Having specified stresses and strains, we next consider the relationship between them. For small displacements, it is a consequence of Hooke's Law that the stresses are proportional to the strains. For the most unsymmetrical medium, this proportionality can be written in the form

$$\begin{aligned}
T_1 &= c_{11}S_1 + c_{12}S_2 + c_{13}S_3 + c_{14}S_4 + c_{15}S_5 + c_{16}S_6 \\
T_2 &= c_{21}S_1 + c_{22}S_2 + c_{23}S_3 + c_{24}S_4 + c_{25}S_6 + c_{26}S_6 \\
T_3 &= c_{31}S_1 + c_{32}S_2 + c_{33}S_3 + c_{34}S_4 + c_{35}S_5 + c_{36}S_6 \\
T_4 &= c_{41}S_1 + c_{42}S_2 + c_{43}S_3 + c_{44}S_4 + c_{45}S_5 + c_{46}S_6 \\
T_5 &= c_{51}S_1 + c_{52}S_2 + c_{53}S_3 + c_{54}S_4 + c_{55}S_5 + c_{56}S_6 \\
T_6 &= c_{61}S_1 + c_{62}S_2 + c_{63}S_3 + c_{64}S_4 + c_{65}S_5 + c_{66}S_6
\end{aligned} \tag{22}$$

where c_{11} for example is an elastic constant expressing the proportionality between the S_1 strain and the T_1 stress in the absence of any other strains.

It can be shown that the law of conservation of energy, it is a necessary consequence that

$$c_{12} = c_{21} \text{ and in general } c_{ij} = c_{ji}. \tag{23}$$

This reduces the number of independent elastic constants for the most unsymmetrical medium to 21. As shown in a later section, any symmetry existing in the crystal will reduce the possible number of elastic constants and simplify the stress strain relationship of equation (22).

Introducing the values of the stresses from (22) in the expression for the potential energy (21), this can be written in the form

$$\begin{aligned}
2PE = {} & c_{11}S_1^2 + 2c_{12}S_1S_2 + 2c_{13}S_1S_3 + 2c_{14}S_1S_4 + 2c_{15}S_1S_5 + 2c_{16}S_1S_6 \\
& + c_{22}S_2^2 + 2c_{23}S_2S_3 + 2c_{24}S_2S_4 + 2c_{25}S_2S_5 + 2c_{26}S_2S_6 \\
& + c_{33}S_3^2 + 2c_{34}S_3S_4 + 2c_{35}S_3S_6 + 2c_{36}S_5S_6 \\
& + c_{44}S_4^2 + 2c_{45}S_4S_5 + 2c_{46}S_4S_6 \\
& + c_{55}S_5^2 + 2c_{56}S_5S_6 \\
& + c_{66}S_6^2.
\end{aligned} \tag{24}$$

The relations (22) thus can be obtained by differentiating the potential energy according to the relation

$$T_1 = \frac{\partial PE}{\partial S_1}; \qquad \cdots; \qquad T_6 = \frac{\partial PE}{\partial S_6}. \tag{25}$$

It is sometimes advantageous to express the strains in terms of the stresses. This can be done by solving the equations (22) simultaneously for the strains resulting in the equations

$$
\begin{aligned}
S_1 &= s_{11}T_1 + s_{12}T_2 + s_{13}T_3 + s_{14}T_4 + s_{15}T_5 + s_{16}T_6 \\
S_2 &= s_{21}T_1 + s_{22}T_2 + s_{23}T_3 + s_{24}T_4 + s_{25}T_5 + s_{26}T_6 \\
S_3 &= s_{31}T_1 + s_{32}T_2 + s_{33}T_3 + s_{34}T_4 + s_{36}T_5 + s_{36}T_6 \\
S_4 &= s_{41}T_1 + s_{42}T_2 + s_{43}T_3 + s_{44}T_4 + s_{45}T_5 + s_{46}T_6 \\
S_5 &= s_{51}T_1 + s_{52}T_2 + s_{53}T_3 + s_{54}T_4 + s_{55}T_5 + s_{56}T_6 \\
S_6 &= s_{61}T_1 + s_{62}T_2 + s_{63}T_3 + s_{64}T_4 + s_{65}T_5 + s_{66}T_6
\end{aligned}
\tag{26}
$$

where

$$
s_{ij} = \frac{(-1)^{i+j}\Delta^c_{ij}}{\Delta^c} \tag{27}
$$

for which Δ^c is the determinant of the c_{ij} terms of (28) and Δ^c_{ij} the minor obtained by suppressing the ith and jth column

$$
\Delta^c = \begin{vmatrix}
c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\
c_{12} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\
c_{13} & c_{23} & c_{33} & c_{34} & c_{35} & c_{36} \\
c_{14} & c_{24} & c_{34} & c_{44} & c_{45} & c_{46} \\
c_{15} & c_{25} & c_{35} & c_{45} & c_{55} & c_{56} \\
c_{16} & c_{26} & c_{36} & c_{46} & c_{5\epsilon} & c_{66}
\end{vmatrix} \tag{28}
$$

Since $c_{ij} = c_{ji}$ it follows that $s_{ij} = s_{ji}$. The potential energy can be expressed in the form.

$$
\begin{aligned}
2PE = {} & s_{11}T_1^2 + 2s_{12}T_1T_2 + 2s_{13}T_1T_3 + 2s_{14}T_1T_4 + 2s_{15}T_1T_5 + 2s_{16}T_1T_6 \\
& + s_{22}T_2^2 + 2s_{23}T_2T_3 + 2s_{24}T_2T_4 + 2s_{25}T_2T_5 + 2s_{26}T_2T_6 \\
& + s_{33}T_3^2 + 2s_{34}T_3T_4 + 2s_{35}T_3T_5 + 2s_{36}T_3T_6 \\
& + s_{44}T_4^2 + 2s_{45}T_4T_5 + 2s_{46}T_4T_6 \\
& + s_{55}T_5^2 + 2s_{56}T_5T_6 \\
& + s_{66}T_6^2.
\end{aligned}
\tag{29}
$$

The relations (26) can then be derived from expressions of the type

$$
S_1 = \frac{\partial PE}{\partial T_1}; \qquad \cdots; \qquad S_6 = \frac{\partial PE}{\partial T_6}. \tag{30}
$$

1.4 *Isothermal and Adiabatic Elastic Constants*

We have so far considered only the elastic relations that can be measured statically at a constant temperature. The elastic constants are then the isothermal constants. For a rapidly vibrating body, however, there is no

chance for heat to equalize and consequently the elastic constants operative are the adiabatic constants determined by the fact that no heat is added or subtracted from any elemental volume. For gases there is a marked difference between the adiabatic and the isothermal constants, but for piezoelectric crystals the difference is small and can usually be neglected.

To investigate the relation existing we can write from the first and second laws of thermodynamics, the relations

$$dU = [T_1\,dS_1 + T_2\,dS_2 + T_3\,dS_3 + T_4\,dS_4 + T_5\,dS_5 + T_6\,dS_6] + \Theta\,d\sigma \tag{31}$$

which expresses the fact that the change in the total energy U is equal to the change in the potential energy plus the added heat energy $dQ = \Theta\,d\sigma$ where Θ is the temperature and σ the entropy. Developing the strains and entropy in terms of the partial differentials of the stresses and temperature, we have

$$\begin{aligned}
dS_1 &= \frac{\partial S_1}{\partial T_1}dT_1 + \frac{\partial S_1}{\partial T_2}dT_2 + \frac{\partial S_1}{\partial T_3}dT_3 + \frac{\partial S_1}{\partial T_4}dT_4 + \frac{\partial S_1}{\partial T_5}dT_5 + \frac{\partial S_1}{\partial T_6}dT_6 + \frac{\partial S_1}{\partial \Theta}d\Theta \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
dS_6 &= \frac{\partial S_6}{\partial T_1}dT_1 + \frac{\partial S_6}{\partial T_2}dT_2 + \frac{\partial S_6}{\partial T_3}dT_3 + \frac{\partial S_6}{\partial T_4}dT_4 + \frac{\partial S_6}{\partial T_5}dT_5 + \frac{\partial S_6}{\partial T_6}dT_6 + \frac{\partial S_6}{\partial \Theta}d\Theta \\
d\sigma &= \frac{\partial \sigma}{\partial T_1}dT_1 + \frac{\partial \sigma}{\partial T_2}dT_2 + \frac{\partial \sigma}{\partial T_3}dT_3 + \frac{\partial \sigma}{\partial T_4}dT_4 + \frac{\partial \sigma}{\partial T_5}dT_5 + \frac{\partial \sigma}{\partial T_6}dT_6 + \frac{\partial \sigma}{\partial \Theta}d\Theta.
\end{aligned} \tag{32}$$

The partial derivatives of the strains with regard to the stresses are readily seen to be the isothermal elastic compliances. The partial derivatives of the strains by the temperatures are the six temperature coefficients of expansion, or

$$\frac{\partial S_1}{\partial \Theta} = \alpha_1\,; \qquad \cdots\,; \qquad \frac{\partial S_6}{\partial \Theta} = \alpha_6\,. \tag{33}$$

To evaluate the partial derivatives of the entropy with respect to the stresses we make use of the fact that U is a perfect differential so that

$$\frac{\partial S_1}{\partial \Theta} = \frac{\partial \sigma}{\partial T_1} = \alpha_1\,; \qquad \cdots\,; \qquad \frac{\partial S_6}{\partial \Theta} = \frac{\partial \sigma}{\partial T_6} = \alpha_6\,. \tag{34}$$

Finally multiplying through the last of equation (32) by Θ we can write them as

$$\begin{aligned} S_1 &= s_{11}^{\Theta}T_1 + s_{12}^{\Theta}T_2 + s_{13}^{\Theta}T_3 + s_{14}^{\Theta}T_4 + s_{15}^{\Theta}T_5 + s_{16}^{\Theta}T_6 + \alpha_1 d\Theta \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ S_6 &= s_{16}^{\Theta}T_1 + s_{26}^{\Theta}T_2 + s_{36}^{\Theta}T_3 + s_{46}^{\Theta}T_4 + s_{56}^{\Theta}T_5 + s_{66}^{\Theta}T_6 + \alpha_6 d\Theta \\ dQ &= \Theta\, d\sigma = \Theta[\alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_3 + \alpha_4 T_4 + \alpha_5 T_5 + \alpha_6 T_6] + \rho C_p d\Theta \end{aligned} \tag{35}$$

since $\Theta \dfrac{\partial \sigma}{\partial \Theta}$ is the total heat capacity of the unit volume at constant stress, which is equal to ρC_p, where ρ is the density and C_p the heat capacity at constant stress per gram of the material.

To get the adiabatic elastic constants which correspond to no heat loss from the element, or $dQ = 0$, $d\Theta$ can be eliminated from (35) giving

$$\begin{aligned} S_1 &= s_{11}^{\sigma}T_1 + s_{12}^{\sigma}T_2 + s_{13}^{\sigma}T_3 + s_{14}^{\sigma}T_4 + s_{15}^{\sigma}T_5 + s_{16}^{\sigma}T_6 + (\alpha_1/\rho C_p)\, dQ \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ S_6 &= s_{16}^{\sigma}T_1 + s_{26}^{\sigma}T_2 + s_{36}^{\sigma}T_3 + s_{46}^{\sigma}T_4 + s_{56}^{\sigma}T_5 + s_{66}^{\sigma}T_6 + (\alpha_6/\rho C_p)\, dQ \end{aligned} \tag{36}$$

where

$$s_{ij}^{\sigma} = s_{ij}^{\Theta} - \frac{\alpha_i \alpha_j \Theta}{\rho C_p}. \tag{37}$$

For example for quartz, the expansion coefficients are

$$\alpha_1 = 14.3 \times 10^{-6}/°\mathrm{C}; \quad \alpha_2 = 14.3 \times 10^{-6}/°\mathrm{C}; \quad \alpha_3 = 7.8 \times 10^{-6}/°\mathrm{C};$$

$$\alpha_4 = \alpha_5 = \alpha_6 = 0$$

The density and specific heat at constant pressure are

$$\rho = 2.65\ \mathrm{grams/cm^3}; \quad C_p = 7.37 \times 10^6\ \mathrm{ergs/cm^3}.$$

Hence the only constants that differ for adiabatic and isothermal values are

$$s_{11} = s_{22};\ s_{12};\ s_{13};\ s_{33}.$$

Taking these values as[3]

$$s_{11}^{\sigma} = 127.9 \times 10^{-14}\ \mathrm{cm^2/dyne}; \quad s_{12}^{\sigma} = -15.35 \times 10^{-14};$$

$$s_{13}^{\sigma} = 11.0 \times 10^{-14}; \quad s_{33}^{\sigma} = 95.6 \times 10^{-14}.$$

We find that the corresponding isothermal values are

$$s_{11}^{\Theta} = 128.2 \times 10^{-14}; \quad s_{12}^{\Theta} = -15.04 \times 10^{-14};$$

$$s_{13}^{\Theta} = 10.83 \times 10^{-14}; \quad s_{33}^{\Theta} = 95.7 \times 10^{-14}\ \mathrm{cm^2/dyne}$$

[3] See "Quartz Crystal Applications" Bell System Technical Journal, Vol. XXII, No. 2, July 1943, W. P. Mason.

at 25°C. or 298° absolute. These differences are probably smaller than the accuracy of the measured constants.

If we express the stresses in terms of the strains by solving equation (35) simultaneously, we find for the stresses

$$\begin{aligned} T_1 &= c_{11}^{\Theta} S_1 + c_{12}^{\Theta} S_2 + c_{13}^{\Theta} S_3 + c_{14}^{\Theta} S_4 + c_{15}^{\Theta} S_5 + c_{16}^{\Theta} S_6 - \lambda_1 \, d\Theta \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ T_6 &= c_{16}^{\Theta} S_1 + c_{26}^{\Theta} S_2 + c_{36}^{\Theta} S_3 + c_{46}^{\Theta} S_4 + c_{56}^{\Theta} S_5 + c_{66}^{\Theta} S_6 - \lambda_6 \, d\Theta \end{aligned} \tag{38}$$

where

$$\begin{aligned} \lambda_1 &= \alpha_1 c_{11}^{\Theta} + \alpha_2 c_{12}^{\Theta} + \alpha_3 c_{13}^{\Theta} + \alpha_4 c_{14}^{\Theta} + \alpha_5 c_{15}^{\Theta} + \alpha_6 c_{16}^{\Theta} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_6 &= \alpha_1 c_{16}^{\Theta} + \alpha_2 c_{26}^{\Theta} + \alpha_3 c_{36}^{\Theta} + \alpha_4 c_{46}^{\Theta} + \alpha_5 c_{56}^{\Theta} + \alpha_6 c_{66}^{\Theta}. \end{aligned}$$

The λ's represent the temperature coefficients of stress when all the strains are zero. The negative sign indicates that a negative stress (a compression) has to be applied to keep the strains zero. If we substitute equations (38) in the last of equations (35), the relation between increments of heat and temperature, we have

$$\begin{aligned} dQ = \Theta d\sigma = \Theta[\lambda_1 S_1 + \lambda_2 S_2 + \lambda_3 S_3 + \lambda_4 S_4 + \lambda_5 S_5 + \lambda_6 S_6] \\ + [\rho C_p - \Theta(\alpha_1\lambda_1 + \alpha_2\lambda_2 + \alpha_3\lambda_3 + \alpha_4\lambda_4 + \alpha_5\lambda_5 + \alpha_6\lambda_6)] d\Theta. \end{aligned} \tag{39}$$

If we set the strains equal to zero, the size of the element does not change, and hence the ratio between dQ and $d\Theta$ should equal ρ times the specific heat at constant volume C_v. We have therefore the relation

$$\rho[C_p - C_v] = \Theta[\alpha_1\lambda_1 + \alpha_2\lambda_2 + \alpha_3\lambda_3 + \alpha_4\lambda_4 + \alpha_5\lambda_5 + \alpha_6\lambda_6]. \tag{40}$$

The relation between the adiabatic and isothermal elastic constants c_{ij} thus becomes

$$c_{ij}^{\sigma} = c_{ij}^{\Theta} + \frac{\lambda_i \lambda_j \Theta}{\rho C_v}. \tag{41}$$

Since the difference between the adiabatic and isothermal constants is so small, no differentiation will be made between them in the following sections.

2. Expression for The Elastic, Piezoelectric, Pyroelectric and Dielectric Relations of a Piezoelectric Crystal

When a crystal is piezoelectric, a potential energy is stored in the crystal when a voltage is applied to the crystal. Hence the energy expressions of (31) requires additional terms to represent the increment of energy dU. If we employ CGS units which have so far been most widely used, as applied

to piezoelectric crystals, the energy stored in any unit volume of the crystal is

$$dU = T_1\,dS_1 + T_2\,dS_2 + T_3\,dS_3 + T_4\,dS_4 + T_5\,dS_5 + T_6\,dS_6 + E_1\frac{dD_1}{4\pi} + E_2\frac{dD_2}{4\pi} + E_3\frac{dD_3}{4\pi} + \Theta\,d\sigma \tag{42}$$

where E_1, E_2 and E_3 are the components of the field existing in the crystal and D_1, D_2 and D_3 the components of the electric displacement. In order to avoid using the factor $1/4\pi$ we make the substitution

$$\frac{D}{4\pi} = \delta. \tag{43}$$

The normal component of δ at any bounding surface is δ_0 the surface charge. On the other hand if we employ the MKS systems of units the energy of any component is given by $E_n dD_n$ directly and in the following formulation δ can be replaced by D.

There are two logical methods of writing the elastic, piezoelectric, pyroelectric and dielectric relations. One considers the independent variables as the stresses, fields, and temperature, and the dependent variables as the strains, displacements and entropy. The other system considers the strains, displacements and entropy as the fundamental independent variables and the stresses, fields, and temperature as the independent variables. The first system appears to be more fundamental for ferroelectric types of crystals.

If we develop the stresses, fields, and temperature in terms of their partial derivatives, we can write

$$T_1 = \left(\frac{\partial T_1}{\partial S_1}\right)_{D,\sigma} dS_1 + \left(\frac{\partial T_1}{\partial S_2}\right)_{D,\sigma} dS_2 + \left(\frac{\partial T_1}{\partial S_3}\right)_{D,\sigma} dS_3 + \left(\frac{\partial T_1}{\partial S_4}\right)_{D,\sigma} dS_4 + \left(\frac{\partial T_1}{\partial S_5}\right)_{D,\sigma} dS_5 + \left(\frac{\partial T_1}{\partial S_6}\right)_{D,\sigma} dS_6 + \left(\frac{\partial T_1}{\partial \delta_1}\right)_{S,\sigma} d\delta_1 + \left(\frac{\partial T_1}{\partial \delta_2}\right)_{S,\sigma} d\delta_2 + \left(\frac{\partial T_1}{\partial \delta_3}\right)_{S,\sigma} d\delta_3 + \left(\frac{\partial T_1}{\partial \sigma}\right)_{S,D} d\delta \tag{44 A}$$

$$T_6 = \left(\frac{\partial T_6}{\partial S_1}\right)_{D,\sigma} dS_1 + \left(\frac{\partial T_6}{\partial S_2}\right)_{D,\sigma} dS_2 + \left(\frac{\partial T_6}{\partial S_3}\right)_{D,\sigma} dS_3 + \left(\frac{\partial T_6}{\partial S_4}\right)_{D,\sigma} dS_4 + \left(\frac{\partial T_6}{\partial S_5}\right)_{D,\sigma} dS_5 + \left(\frac{\partial T_6}{\partial S_6}\right)_{D,\sigma} dS_6 + \left(\frac{\partial T_6}{\partial \delta_1}\right)_{S,\sigma} d\delta_1 + \left(\frac{\partial T_6}{\partial \delta_2}\right)_{S,\sigma} d\delta_2 + \left(\frac{\partial T_6}{\partial \delta_3}\right)_{S,\sigma} d\delta_3 + \left(\frac{\partial T_6}{\partial \sigma}\right)_{S,D} d\sigma$$

$$E_z = E_1 = \frac{\partial E_1}{\partial S_1}\Big)_{D,\sigma} dS_1 + \frac{\partial E_1}{\partial S_2}\Big)_{D,\sigma} dS_2 + \frac{\partial E_1}{\partial S_3}\Big)_{D,\sigma} dS_3 + \frac{\partial E_1}{\partial S_4}\Big)_{D,\sigma} dS_4$$
$$+ \frac{\partial E_1}{\partial S_5}\Big)_{D,\sigma} dS_5 + \frac{\partial E_1}{\partial S_6}\Big)_{D,\sigma} dS_6 + \frac{\partial E_1}{\partial \delta_1}\Big)_{S,\sigma} d\delta_1 + \frac{\partial E_1}{\partial \delta_2}\Big)_{S,\sigma} d\delta_2$$
$$+ \frac{\partial E_1}{\partial \delta_3}\Big)_{S,\sigma} d\delta_3 + \frac{\partial E_1}{\partial \sigma}\Big)_{S,D} d\sigma$$

. . . . . .

$$E_z = E_3 = \frac{\partial E_3}{\partial S_1}\Big)_{D,\sigma} dS_1 + \frac{\partial E_3}{\partial S_2}\Big)_{D,\sigma} dS_2 + \frac{\partial E_3}{\partial S_3}\Big)_{D,\sigma} dS_3 + \frac{\partial E_3}{\partial S_4}\Big)_{D,\sigma} dS_4$$
$$+ \frac{\partial E_3}{\partial S_5}\Big)_{D,\sigma} dS_5 + \frac{\partial E_3}{\partial S_6}\Big)_{D,\sigma} dS_6 + \frac{\partial E_3}{\partial \delta_1}\Big)_{S,\sigma} d\delta_1 + \frac{\partial E_3}{\partial \delta_2}\Big)_{S,\sigma} d\delta_2$$
$$+ \frac{\partial E_3}{\partial \delta_3}\Big)_{S,\sigma} d\delta_3 + \frac{\partial E_3}{\partial \sigma}\Big)_{S,D} d\sigma \qquad (44\ \text{B})$$

$$d\Theta = \frac{\partial \Theta}{\partial S_1}\Big)_{D,\sigma} dS_1 + \frac{\partial \Theta}{\partial S_2}\Big)_{D,\sigma} dS_2 + \frac{\partial \Theta}{\partial S_3}\Big)_{D,\sigma} dS_3 + \frac{\partial \Theta}{\partial S_4}\Big)_{D,\sigma} dS_4$$
$$+ \frac{\partial \Theta}{\partial S_5}\Big)_{D,\sigma} dS_5 + \frac{\partial \Theta}{\partial S_6}\Big)_{D,\sigma} dS_6 + \frac{\partial \Theta}{\partial \delta_1}\Big)_{S,\sigma} d\delta_1 + \frac{\partial \Theta}{\partial \delta_2}\Big)_{S,\sigma} d\delta_2$$
$$+ \frac{\partial \Theta}{\partial \delta_3}\Big)_{S,\sigma} d\delta_3 + \frac{\partial \Theta}{\partial \sigma}\Big)_{S,D} d\sigma.$$

The subscripts under the partial derivatives indicate the quantities kept constant. A subscript D indicates that the electric induction is held constant, a subscript σ indicates that the entropy is held constant, while a subscript S indicates that the strains are held constant.

Examining the first equation, we see that the partial derivatives of the stress T_1 by the strains are the elastic constants c_{ij} which determine the ratios between the stress T_1 and the appropriate strain with all other strains equal to zero. To indicate the conditions for the partial derivatives, the superscripts D and σ are given to the elastic constants and they are written $c_{ij}^{D,\sigma}$. The partial derivatives of the stresses by $\delta = D/4\pi$ are the piezoelectric constants h_{ij} which measure the increases in stress necessary to hold the crystal free from strain in the presence of a displacement. Since if the crystal tends to expand on the application of a displacement, the stress to keep it from expanding has to be a compression or negative stress, the negative sign is given to the h_{ij}^{σ} constants. As the only meaning of the h constants is obtained by measuring the ratio of the stress to $\delta = D/4\pi$ at constant strains, no superscript S is added. However there is a difference between isothermal and adiabatic piezoelectric constants in general, so

that these piezoelectric constants are written h^{σ}_{ij}. Finally the last partial derivatives of the stresses by the entropy σ can be written

$$\frac{\partial T_n}{\partial \sigma}\Big)_{S,D} d\sigma = \frac{1}{\Theta}\frac{\partial T_n}{\partial \sigma}\Big)_{S,D} \Theta\, d\sigma = \frac{1}{\Theta}\frac{\partial T_n}{\partial \sigma}\Big)_{S,D} dQ = -\gamma_n^{S,D}\, dQ \qquad (45)$$

where dQ is the added heat. We designate $1/\Theta$ times the partial derivative as $-\gamma_n^{S,D}$ and note that it determines the negative stress (compression) necessary to put on the crystal to keep it from expanding when an increment of heat dQ is added to the crystal. The electric displacement is held constant and hence the superscripts S, and D are used. The first six equations then can be written in the form

$$\begin{aligned} T_n = {} & c_{n1}^{D,\sigma} S_1 + c_{n2}^{D,\sigma} S_2 + c_{n3}^{D,\sigma} S_3 + c_{n4}^{D,\sigma} S_4 + c_{n5}^{D,\sigma} S_5 + c_{n6}^{D,\sigma} S_6 \\ & - h_{n1}^{\sigma}\delta_1 - h_{n2}^{\sigma}\delta_2 - h_{n3}^{\sigma}\delta_3 - \gamma_n^{S,D}\, dQ. \end{aligned} \qquad (46)$$

To evaluate the next three equations involving the fields, we make use of the fact that the expression for dU in equation (42) is a perfect differential. As a consequence there are relations between the partial derivatives, namely

$$\frac{\partial T_m}{\partial \delta_n} = \frac{\partial E_n}{\partial S_m}; \qquad \frac{\partial T_m}{\partial \sigma} = \frac{\partial \Theta}{\partial S_m}; \qquad \frac{\partial E_n}{\partial \sigma} = \frac{\partial \Theta}{\partial \delta_n}. \qquad (47)$$

We note also that

$$\frac{\partial E_m}{\partial \delta_n}\Big)_{S,\sigma} = 4\pi\beta_{mn}^{S,\sigma} \qquad (48)$$

where β is the so called "impermeability" matrix obtained from the dielectric matrix ϵ_{nm} by means of the equation

$$\beta_{mn} = \frac{(-1)^{m+n}\Delta^{m,n}}{\Delta} \qquad (49)$$

where Δ is the determinant

$$\Delta = \begin{vmatrix} \epsilon_{11}, & \epsilon_{12}, & \epsilon_{13} \\ \epsilon_{12}, & \epsilon_{22}, & \epsilon_{23} \\ \epsilon_{13}, & \epsilon_{23}, & \epsilon_{33} \end{vmatrix} \qquad (50)$$

and $\Delta^{m,n}$ the minor obtained by suppressing the mth row and nth column.

The partial derivatives of the fields by the entropy can be written

$$\frac{\partial E_m}{\partial \sigma}\Big)_{S,D} d\sigma = \frac{1}{\Theta}\frac{\partial E_m}{\partial \sigma}\Big)_{S,D} \Theta\, \delta\sigma = \frac{1}{\Theta}\frac{\partial E_m}{\partial \sigma}\Big)_{S,D} dQ = -q_m^{S,D}\, dQ \qquad (51)$$

where $q_n^{S,D}$ is a pyroelectric constant measuring the increase in field required to produce a zero charge on the surface when a heat dQ is added to the

crystal. Since the voltage will be of opposite sign to the charge generated on the surface of the crystal in the absence of this counter voltage a negative sign is given to $q_n^{S,D}$.

Finally the last partial derivative

$$\frac{\partial\Theta}{\partial\sigma}\Big)_{S,D} d\sigma = \frac{1}{\Theta}\frac{\partial\Theta}{\partial\sigma}\Big)_{S,D} \Theta\, d\sigma = \frac{1}{\Theta}\frac{\partial\Theta}{\partial\sigma}\Big)_{S,D} dQ = \frac{dQ}{\rho C_v^D} \tag{52}$$

represents the ratio of the increase in temperature due to the added amount of heat dQ when the strains and electric displacements are held constant. It is therefore the inverse of the specific heat at constant volume and constant electric displacement per gram of material times the density ρ. Hence the ten equations of equation (44) can be written in the generalized forms

$$\begin{aligned}
T_n &= c_{n1}^{D,\sigma}S_1 + c_{n2}^{D,\sigma}S_2 + c_{n3}^{D,\sigma}S_3 + c_{n4}^{D,\sigma}S_4 + c_{n5}^{D,\sigma}S_5 + c_{n6}^{D,\sigma}S_6 \\
&\quad - h_{n1}^{\sigma}\delta_1 - h_{n2}^{\sigma}\delta_2 - h_{n3}^{\sigma}\delta_3 - \gamma_n^{S,D}\, dQ \\
E_m &= -h_{1m}^{\sigma}S_1 - h_{2m}^{\sigma}S_2 - h_{3m}^{\sigma}S_3 - h_{4m}^{\sigma}S_4 - h_{5m}^{\sigma}S_5 - h_{6m}^{\sigma}S_6 + 4\pi\beta_{m1}^{S,\sigma}\delta_1 \\
&\quad + 4\pi\beta_{m2}^{S,\sigma}\delta_2 + 4\pi\beta_{m3}^{S,\sigma}\delta_3 - q_m^{S,D}\, dQ \qquad (53) \\
d\Theta &= -\Theta[\gamma_1^{S,D}S_1 + \gamma_2^{S,D}S_2 + \gamma_3^{S,D}S_3 + \gamma_4^{S,D}S_4 + \gamma_5^{S,D}S_5 + \gamma_6^{S,D}S_6] \\
&\quad - \Theta[q_1^{S,D}\delta_1 + q_2^{S,D}\delta_2 + q_3^{S,D}\delta_3] + \frac{dQ}{\rho C_v^D}.
\end{aligned}$$

$$n = 1 \text{ to } 6;\ m = 1 \text{ to } 3$$

If, as is usually the case with vibrating crystals the vibration occurs with no interchange of heat between adjacent elements $dQ = 0$ and the ten equations reduce to the usual nine given by the general forms

$$\begin{aligned}
T_n &= c_{n1}^{D}S_1 + c_{n2}^{D}S_2 + c_{n3}^{D}S_3 + c_{n4}^{D}S_4 + c_{n5}^{D}S_5 + c_{n6}^{D}S_6 \\
&\quad - h_{n1}\delta_1 - h_{n2}\delta_2 - h_{n3}\delta_3 \\
E_m &= -h_{1m}S_1 - h_{2m}S_2 - h_{3m}S_3 - h_{4m}S_4 - h_{5m}S_5 - h_{6m}S_6 \\
&\quad + 4\pi\beta_{m1}^{S}\delta_1 + 4\pi\beta_{m2}^{S}\delta_2 + 4\pi\beta_{m3}^{S}\delta_3.
\end{aligned} \tag{54}$$

In these equations the superscript σ has been dropped since the ordinary constants are adiabatic. The tenth equation of (53) determines the increase in temperature caused by the strains and displacements in the absence of any flow of heat.

If we introduce the expression of equations (53) into equation (42) the total energy of the crystal is per unit volume.

$$
\begin{aligned}
2U = {} & c_{11}^{D,\sigma} S_1^2 + 2c_{12}^{D,\sigma} S_1 S_2 + 2c_{13}^{D,\sigma} S_1 S_3 + 2c_{14}^{D,\sigma} S_1 S_4 + 2c_{15}^{D,\sigma} S_1 S_5 + 2c_{16}^{D,\sigma} S_1 S_6 \\
& + c_{22}^{D,\sigma} S_2^2 + 2c_{23}^{D,\sigma} S_2 S_3 + 2c_{24}^{D,\sigma} S_2 S_4 + 2c_{25}^{D,\sigma} S_2 S_5 + 2c_{26}^{D,\sigma} S_2 S_6 \\
& + c_{33}^{D,\sigma} S_3^2 + 2c_{34}^{D,\sigma} S_3 S_4 + 2c_{35}^{D,\sigma} S_3 S_5 + 2c_{36}^{D,\sigma} S_3 S_6 \\
& + c_{44}^{D,\sigma} S_4^2 + 2c_{45}^{D,\sigma} S_4 S_5 + 2c_{46}^{D,\sigma} S_4 S_6 \\
& + c_{55}^{D,\sigma} S_5^2 + 2c_{56}^{D,\sigma} S_5 S_6 \\
& + c_{66}^{D,\sigma} S_6^2 \qquad (55)
\end{aligned}
$$

$$
\begin{aligned}
& - (2h_{11}^{\sigma} \delta_1 S_1 + 2h_{12}^{\sigma} \delta_1 S_2 + 2h_{13}^{\sigma} \delta_1 S_3 + 2h_{14}^{\sigma} \delta_1 S_4 + 2h_{15}^{\sigma} \delta_1 S_5 + 2h_{16}^{\sigma} \delta_1 S_6) \\
& - (2h_{21}^{\sigma} \delta_2 S_1 + 2h_{22}^{\sigma} \delta_2 S_2 + 2h_{23}^{\sigma} \delta_2 S_3 + 2h_{24}^{\sigma} \delta_2 S_4 + 2h_{25}^{\sigma} \delta_2 S_5 + 2h_{26}^{\sigma} \delta_2 S_6) \\
& - (2h_{31}^{\sigma} \delta_3 S_1 + 2h_{32}^{\sigma} \delta_3 S_2 + 2h_{33}^{\sigma} \delta_3 S_3 + 2h_{34}^{\sigma} \delta_3 S_4 + 2h_{35}^{\sigma} \delta_3 S_5 + 2h_{36}^{\sigma} \delta_3 S_6) \\
& - (2\gamma_1^{S,D} S_1 dQ + 2\gamma_2^{S,D} S_2 dQ + 2\gamma_3^{S,D} S_3 \, dQ \\
& \qquad + 2\gamma_4^{S,D} S_4 dQ + 2\gamma_5^{S,D} S_5 dQ + 2\gamma_6^{S,D} S_6 dQ) \\
& + 4\pi [\beta_{11}^{S,\sigma} \delta_1^2 + 2\beta_{12}^{S,\sigma} \delta_1 \delta_2 + 2\beta_{13}^{S,\sigma} \delta_1 \delta_3 + \beta_{22}^{S,\sigma} \delta_2^2 + 2\beta_{23}^{S,\sigma} \delta_2 \delta_3 + \beta_{33}^{S,\sigma} \delta_3^2] \\
& - (2q_1^{S,D} \delta_1 dQ + 2q_2^{S,D} \delta_2 dQ + 2q_3^{S,D} \delta_3 dQ) + \frac{(dQ)^2}{{}_{\rho} C_v^D} .
\end{aligned}
$$

Equations (53) can be derived from this expression by employing the partial derivatives

$$
T_n = \frac{\partial U}{\partial S_n} \; ; \quad E_m = \frac{\partial U}{\partial \delta_m} \; ; \quad d\Theta = \frac{\partial U}{\partial (dQ)} \qquad (56)
$$

The other form for writing the elastic, piezoelectric, pyroelectric and dielectric relations is to take the strains, displacements, and entropy as the fundamental variables and the stresses, fields and temperature increments as the dependent variables. If we develop them in terms of their partial derivatives as was done in (44), use the relations between the partial derivatives shown in equation (57).

$$
\frac{\partial \delta_m}{\partial T_n} = \frac{\partial S_n}{\partial E_m} \; ; \quad \frac{\partial S_n}{\partial \Theta} = \frac{\partial \sigma}{\partial T_n} \; ; \quad \frac{\partial \delta_m}{\partial \Theta} = \frac{\partial \sigma}{\partial E_m} \qquad (57)
$$

and substitute for the partial derivatives their equivalent elastic, piezoelectric, pyroelectric, temperature expansions, dielectric and specific heat constants, there are 10 equations of the form

$$S_n = s^{E,\Theta}_{n1}T_1 + s^{E,\Theta}_{n2}T_2 + s^{E,\Theta}_{n3}T_3 + s^{E,\Theta}_{n4}T_4 + s^{E,\Theta}_{n5}T_5 + s^{E,\Theta}_{n6}T_6 + d^{\Theta}_{n1}E_1 + d^{\Theta}_{n2}E_2 + d^{\Theta}_{n3}E_3 + \alpha^{E}_{n}d\Theta$$

$$\delta_m = d^{\Theta}_{1m}T_1 + d^{\Theta}_{2m}T_2 + d^{\Theta}_{3m}T_3 + d^{\Theta}_{4m}T_4 + d^{\Theta}_{5m}T_5 + d^{\Theta}_{6m}T_6 + \frac{\epsilon^{T,\Theta}_{m1}}{4\pi}E_1 + \frac{\epsilon^{T,\Theta}_{m2}}{4\pi}E_2 + \frac{\epsilon^{T,\Theta}_{m3}}{4\pi}E_3 + p^{T}_{m}d\Theta \qquad (58)$$

$$dQ = \Theta\, d\sigma = \Theta[\alpha^{E}_1T_1 + \alpha^{E}_2T_2 + \alpha^{E}_3T_3 + \alpha^{E}_4T_4 + \alpha^{E}_5T_5 + \alpha^{E}_6T_6] + \Theta[p^{T}_1E_1 + p^{T}_2E_2 + p^{T}_3E_3] + pC^{E}_{p}d\Theta.$$

$$n = 1 \text{ to } 6, \qquad m = 1 \text{ to } 3$$

The superscripts E, Θ, and T indicate respectively constant field, constant temperature and constant stress for the measurements of the respective constants. It will be noted that the elastic compliance and the piezoelectric constants d_{mn} are for isothermal conditions. The α^{E} constants are the temperature expansion constants measured at constant field, while the p^{T} constants are the pyroelectric constants relating the ratio of $\delta = D/4\pi$ to increase in temperature $d\Theta$, measured at constant stress. Since there is constant stress, these constants take into account not only the "true" pyroelectric effect which is the ratio of $\delta = D/4\pi$ to the temperature at constant volume, but also the so called "false" pyroelectric effect of the first kind which is the polarization caused by the temperature expansion of the crystal. This appears to be a misnomer. A better designation for the two effects is the pyroelectric effect at constant strain and the pyroelectric effect at constant stress. C^{E}_{p} is the specific heat at constant pressure and constant field.

If we substitute these equations into equation (42), the total free energy becomes

$$2U = \sum_{m=1}^{6}\sum_{n=1}^{6} s^{E,\Theta}_{mn}T_mT_n + 2\sum_{n=1}^{6}\sum_{0=1}^{3} d^{\Theta}_{n0}T_nE_0 + 2\sum_{n=1}^{6}\alpha^{E}_{n}T_nd\Theta + \sum_{0=1}^{3}\sum_{p=1}^{3}\frac{\epsilon^{T,\Theta}_{0p}}{4\pi}E_0E_p + 2\sum_{0=1}^{3}p^{T}_{0}E_pd\Theta + \frac{\rho C^{E}_{p}}{\Theta}d\Theta. \qquad (59)$$

Equation (58) can then be obtained by partial derivatives of the sort

$$S_n = \frac{\partial U}{\partial T_n}; \qquad \delta_0 = \frac{\partial U}{\partial E_p}; \qquad d\sigma = \frac{dQ}{\Theta} = \frac{\partial U}{\partial(d\Theta)}.$$

By tensor transformations the expression for U in (59) can be shown to be equal to the expression for U in (55).

The adiabatic equations holding for a rapidly vibrating crystal can be

obtained by setting dQ equal to zero in the last of equations (58) and eliminating $d\Theta$ from the other nine equations. The resulting equations are

$$\begin{aligned} S_n &= s^E_{n1}T_1 + s^E_{n2}T_2 + s^E_{n3}T_3 + s^E_{n4}T_4 \\ &\quad + s^E_{n5}T_5 + s^E_{n6}T_6 + d_{n1}E_1 + d_{n2}E_2 + d_{n3}E_3 \\ \delta_m &= d_{1m}T_1 + d_{2m}T_2 + d_{3m}T_3 + d_{4m}T_4 \\ &\quad + d_{5m}T_5 + d_{6m}T_6 + \frac{\epsilon^T_{m1}}{4\pi}E_1 + \frac{\epsilon^T_{m2}}{4\pi}E_2 + \frac{\epsilon^T_{m3}}{4\pi}E_3 \end{aligned} \tag{60}$$

where the symbol σ for adiabatic is understood and where the relations between the isothermal and adiabatic constants are given by

$$s^{E,\sigma}_{mn} = s^{E,\Theta}_{mn} - \frac{\alpha^E_m \alpha^E_n \Theta}{\rho C^E_p}; \quad d^\sigma_{\ell} = d^\Theta_{\ell m} - \frac{\alpha^E_m p^T_\ell \Theta}{\rho C^E_p}; \quad \frac{\epsilon^{T,\sigma}_{mn}}{4\pi} = \frac{\epsilon^{T,\Theta}_{mn}}{4\pi} - \frac{p^T_m p^T_n \Theta}{\rho C^E_p}.$$

Hence the piezoelectric and dielectric constants are identical for isothermal and adiabatic conditions provided the crystal is not pyroelectric, but differ if the crystal is pyroelectric. The difference between the adiabatic and isothermal elastic compliances was discussed in section (1.4) and was shown to be small. Hence the equations in the form (60) are generally used in discussing piezoelectric crystals.

Two other mixed forms are also used but a discussion of them will be delayed until a tensor notation for piezoelectric crystals has been discussed. This simplifies the writing of such equations.

3. General Properties of Tensors

The expressions for the piezoelectric relations discussed in section 2 can be considerably abbreviated by expressing them in tensor form. Furthermore, the calculation of elastic constants for rotated crystals is considerably simplified by the geometrical transformation laws established for tensors. Hence it has seemed worthwhile to express the elastic, electric, and piezoelectric relations of a piezoelectric crystal in tensor form. It is the purpose of this section to discuss the general properties of tensors applicable to Cartesian coordinates.

If we have two sets of rectangular axes (Ox, Oy, Oz) and (Ox', Oy', Oz) having the same origin, the coordinates of any point P with respect to the second set are given in terms of the first set by the equations

$$\begin{aligned} x' &= \ell_1 x + m_1 y + n_1 z \\ y' &= \ell_2 x + m_2 y + n_2 z \\ z' &= \ell_3 x + m_3 y + n_3 z. \end{aligned} \tag{61}$$

The quantities $(\ell_1, \cdots, n_3)$ are the cosines of the angles between the various axes; thus ℓ_1 is the cosine of the angle between the axes Ox', and Ox; n_3 the cosine of the angle between Oz' and Oz, and so on. By solving the equations (61) simultaneously, the coordinates x, y, z can be expressed in terms of x', y', z' by the equations.

$$\begin{aligned} x &= \ell_1 x' + \ell_2 y' + \ell_3 z' \\ y &= m_1 x' + m_2 y' + m_3 z' \\ z &= n_1 x' + n_2 y' + n_3 z'. \end{aligned} \tag{62}$$

We can shorten the writing of equations (61) and (62) considerably by changing the notation. Instead of x, y, z let us write x_1, x_2, x_3 and in place of x', y', z' we write x_1', x_2', x_3'. We can now say that the coordinates with respect to the first system are x_i, where i may be 1, 2, 3 while those with respect of the second system are x_j', where $j = 1, 2$ or 3. Then in (61) each coordinate x_j' is expressed as the sum of three terms depending on the three x_i. Each x_i is associated with the cosine of the angle between the direction of x_i increasing and that of x_j' increasing. Let us denote this cosine by a_{ij}. Then we have for all values of j,

$$x_j' = a_{1j}x_1 + a_{2j}x_2 + a_{3j}x_3 = \sum_{i=1}^{3} a_{ij}x_i. \tag{63}$$

Conversely equation (62) can be written

$$x_i = \sum_{j=1}^{3} a_{ij}x_j' \tag{64}$$

where the a_{ij} have the same value as in (63), for the same values of i and j, since in both cases the cosine of the angle is between the values of x_i and x_j increasing. Such a set of three quantities involving a relation between two coordinate systems is called a tensor of the first rank or a vector.

We note that each of the equations (63), (64) is really a set of three equations. Where the suffix i or j appears on the left it is to be given in turn all the values 1, 2, 3 and the resulting equation is one of the set. In each such equation the right side is the sum of three terms obtained by giving j or i the values 1, 2, 3 in turn and adding. Whenever such a summation occurs a suffix is repeated in the expression for the general term as $a_{ij}x_j'$. We make it a regular convention that whenever a suffix is repeated it is to be given all possible values and that the terms are to be added for all. Then (63) can be written simply as

$$x_j' = a_{ij}x_i$$

the summation being automatically understood by the convention.

There are single quantities such as mass and distance, that are the same for all systems of coordinates. These are called tensors of the zero rank or scalars.

Consider now two tensors of the first rank u_i and v_k. Suppose that each component of one is to be multiplied by each component of the other, then we obtain a set of nine quantities expressed by $u_i v_k$, where i and k are independently given all the values 1, 2, 3. The components of $u_i v_k$ with respect to the x'_j set of axes are $u'_j v'_\ell$, and

$$u'_j v'_\ell = (a_{ij} u_i)(a_{k\ell} v_k) = a_{ij} a_{k\ell} u_i v_k . \tag{65}$$

The suffixes i and k are repeated on the right. Hence (65) represents nine equations, each with nine terms. Each term on the right is the product of two factors, one of the form $a_{ij} a_{k\ell}$, depending only on the orientation of the axes, and the other of the form $u_i v_k$, representing the products of the components referred to the original axes. In this way the various $u'_j v'_\ell$ can be obtained in terms of the original $u_i v_k$. But products of vectors are not the only quantities satisfying the rule. In general a set of nine quantities w_{ik} referred to a set of axes, and transformed to another set by the rule

$$w'_{j\ell} = a_{ij} a_{k\ell} w_{ik} \tag{66}$$

is called a tensor of the second rank.

Higher orders tensors can be formed by taking the products of more vectors. Thus a set of n quantities that transforms like the vector product $x_i x_j \cdots x_p$ is called a tensor of rank n, where n is the number of factors.

On the right hand side of (66) the i and k are dummy suffices; that is, they are given the numbers 1 to 3 and summed. It, therefore, makes no difference which we call i and which k so that

$$w'_{j\ell} = a_{ij} a_{k\ell} w_{ik} = a_{kj} a_{i\ell} w_{k\ell} . \tag{67}$$

Hence $w_{k\ell}$ transforms by the same rule as w_{ik} and hence is a tensor of the second rank. The importance of this is that if we have a set of quantities

$$\begin{vmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{vmatrix} \tag{68}$$

which we know to be a tensor of the second rank, the set of quantities

$$\begin{vmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \end{vmatrix} \tag{69}$$

is another tensor of the second rank. Hence the sum $(w_{ik} + w_{ki})$ and the difference $(w_{ik} - w_{ki})$ are also tensors of the second rank. The first of

these has the property that it is unaltered by interchanging i and k and therefore it is called a symmetrical tensor. The second has its components reversed in sign when i and k are interchanged. It is therefore an antisymmetrical tensor. Clearly in an antisymmetric tensor the leading diagonal components will all be zero, i.e., those with $i = k$ will be zero. Now since

$$w_{ik} = \tfrac{1}{2}(w_{ik} + w_{ki}) + \tfrac{1}{2}(w_{ik} - w_{ki}) \tag{70}$$

we can consider any tensor of the second rank as the sum of a symmetrical and an antisymmetrical tensor. Most tensors in the theory of elasticity are symmetrical tensors.

The operation of putting two suffixes in a tensor equal and adding the terms is known as contraction of the tensor. It gives a tensor two ranks lower than the original one. If for instance we contract the tensor $u_i v_k$ we obtain

$$u_i v_i = u_1 v_1 + u_2 v_2 + u_3 v_3 \tag{71}$$

which is the scalar product of u_i and v_k and hence is a tensor of zero rank.

We wish now to derive the formulae for tensor transformation to a new set of axes. For a tensor of the first rank (a vector) this has been given by equation (61). But the direction consines ℓ_1 to n_3 can be expressed in the form

$$\begin{aligned}
\ell_1 &= \frac{\partial x'}{\partial x} = \frac{\partial x_1'}{\partial x_1}; \qquad m_1 = \frac{\partial x'}{\partial y} = \frac{\partial x_1'}{\partial x_2}; \qquad n_1 = \frac{\partial x'}{\partial z} = \frac{\partial x_1'}{\partial x_3} \\
\ell_2 &= \frac{\partial y'}{\partial x} = \frac{\partial x_2'}{\partial x_1}; \qquad m_2 = \frac{\partial y'}{\partial y} = \frac{\partial x_2'}{\partial x_2}; \qquad n_2 = \frac{\partial y'}{\partial z} = \frac{\partial x_2'}{\partial x_3} \\
\ell_3 &= \frac{\partial z'}{\partial x} = \frac{\partial x_3'}{\partial x_1}; \qquad m_3 = \frac{\partial z'}{\partial y} = \frac{\partial x_3'}{\partial x_2}; \qquad n_3 = \frac{\partial z'}{\partial z} = \frac{\partial x_3'}{\partial x_3}
\end{aligned} \tag{72}$$

Hence equation (61) can be expressed in the tensor form

$$x_j' = \frac{\partial x_j'}{\partial x_i} x_i = a_{ij} x_i . \tag{73}$$

Similarly since a tensor of the second rank can be regarded as the product of two vectors, it can be transformed according to the equation

$$x_j' x_\ell' = \left(\frac{\partial x_j'}{\partial x_i} x_i\right)\left(\frac{\partial x_\ell'}{\partial x_k} x_k\right) = \frac{\partial x_j'}{\partial x_i}\frac{\partial x_\ell'}{\partial x_k} x_i x_k \tag{74}$$

which can also be expressed in the generalized form

$$w_{j\ell}' = \frac{\partial x_j'}{\partial x_i}\frac{\partial x_\ell'}{\partial x_k} w_{ik} . \tag{75}$$

In general the transformation equation of a tensor of the nth rank can be written

$$X'_{k_1 \cdots k_n} = \frac{\partial x_{k_2'}}{\partial x_{j_1}} \frac{\partial x_{k_2'}}{\partial x_{j_2}} \cdots \frac{\partial x_{k_n'}}{\partial x_{j_n}} X_{j_1 j_2 \cdots j_n} . \tag{76}$$

4. Application of Tensor Notation to the Elastic, Piezoelectric and Dielectric Equations of a Crystal

Let us consider the stress components of equation (7)

$$\begin{vmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{vmatrix}$$

from which equation (8) is derived

$$T_{xy} = T_{yx} \, ; \, T_{xz} = T_{zx} \, ; \, T_{yz} = T_{zy}$$

and designate them in the manner shown by equation (77) to correspond with tensor notations

$$\begin{vmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{vmatrix} = \begin{vmatrix} T_{11} & T_{12} & T_{13} \\ T_{12} & T_{22} & T_{23} \\ T_{13} & T_{23} & T_{33} \end{vmatrix} \tag{77}$$

by virtue of the relations of (8). We wish to show now that the set of 9 elements of the equation constitutes a tensor, and by virtue of the relations of (8) a symmetrical tensor.

The transformation of the stress components to a new set of axes x', y', z' has been shown by Love[4] to take the form

$$T'_{xx} = \ell_1^2 T_{xx} + m_1^2 T_{yy} + n_1^2 T_{zz} + 2\ell_1 m_1 T_{xy} + 2\ell_1 n_1 T_{xz} + 2m_1 n_1 T_{yz} \tag{78}$$

$$T'_{xy} = \ell_1 \ell_2 T_{xx} + m_1 m_2 T_{yy} + n_1 n_2 T_{zz} + (\ell_1 m_2 + \ell_2 m_1) T_{xy} + (\ell_1 n_2 + \ell_2 n_1) T_{xz} + (m_1 n_2 + n_1 m_2) T_{yz}$$

where ℓ_1 to n_3 are the direction cosines between the axes as specified by equation (61). Noting that from (72)

$$\ell_1 = \frac{\partial x_1'}{\partial x_1}, \qquad \cdots, \qquad n_3 = \frac{\partial x_3'}{\partial x_3}$$

the first of these equations can be put in the form

[4] See "Theory of Elasticity," Love, Page 80.

$$T'_{11} = \left(\frac{\partial x'^2_1}{\partial x_1}\right) T_{11} + \frac{\partial x'_1}{\partial x_1}\frac{\partial x'_1}{\partial x_2} T_{12} + \frac{\partial x'_1}{\partial x_1}\frac{\partial x'_1}{\partial x_3} T_{13}$$

$$+ \frac{\partial x'_1}{\partial x_2}\frac{\partial x'_1}{\partial x_1} T_{21} + \left(\frac{\partial x'^2_1}{\partial x_2}\right) T_{22} + \frac{\partial x'_1}{\partial x_2}\frac{\partial x'_1}{\partial x_3} T_{23} = \frac{\partial x'_1}{\partial x_k}\frac{\partial x'_1}{\partial x_\ell} T_{k\ell} \qquad (79)$$

$$+ \frac{\partial x'_1}{\partial x_3}\frac{\partial x'_1}{\partial x_1} T_{31} + \frac{\partial x'_1}{\partial x_3}\frac{\partial x'_1}{\partial x_2} T_{32} + \left(\frac{\partial x_1}{\partial x_3}\right)^2 T_{33}$$

while the last equation takes the form

$$T'_{12} = \frac{\partial x'_1}{\partial x_1}\frac{\partial x'_2}{\partial x_1} T_{11} + \frac{\partial x'_1}{\partial x_1}\frac{\partial x'_2}{\partial x_2} T_{12} + \frac{\partial x'_1}{\partial x_1}\frac{\partial x'_2}{\partial x_3} T_{13}$$

$$+ \frac{\partial x'_1}{\partial x_2}\frac{\partial x'_2}{\partial x_1} T_{21} + \frac{\partial x'_1}{\partial x_2}\frac{\partial x'_2}{\partial x_2} T_{22} + \frac{\partial x'_1}{\partial x_2}\frac{\partial x'_2}{\partial x_3} T_{23} = \frac{\partial x'_1}{\partial x_k}\frac{\partial x'_2}{\partial x_\ell} T_{k\ell} \qquad (80)$$

$$+ \frac{\partial x'_1}{\partial x_3}\frac{\partial x'_2}{\partial x_1} T_{31} + \frac{\partial x'_1}{\partial x_3}\frac{\partial x'_2}{\partial x_2} T_{32} + \frac{\partial x'_1}{\partial x_3}\frac{\partial x'_2}{\partial x_3} T_{33}.$$

The general expression for any component then is

$$T'_{ij} = \frac{\partial x'_i}{\partial x_k}\frac{\partial x'_j}{\partial x_\ell} T_{k\ell} \qquad (81)$$

which is the transformation equation of a tensor of the second rank. Hence the stress components satisfy the conditions for a second rank tensor.

The strain components

$$\begin{vmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{vmatrix}$$

do not however satisfy the conditions for a second rank tensor. This is shown by the transformation of strain components to a new set of axes, which have been shown by Love to satisfy the equations

$$S'_{xx} = \ell_1^2 S_{xx} + m_1^2 S_{yy} + n_1^2 S_{zz} + \ell_1 m_1 S_{xy} + \ell_1 n_1 S_{zx} + m_1 n_1 S_{yz}$$

$$\cdots\cdots\cdots

If, however, we take the strain components as

$$S_{11} = S_{xx} = \frac{\partial \xi}{\partial x}; \qquad S_{22} = S_{yy} = \frac{\partial \eta}{\partial y}; \qquad S_{33} = S_{zz} = \frac{\partial \zeta}{\partial z}$$

$$S_{12} = S_{21} = \frac{S_{xy}}{2} = \frac{1}{2}\left(\frac{\partial \eta}{\partial x} + \frac{\partial \xi}{\partial y}\right); \qquad S_{13} = S_{31} = \frac{S_{zx}}{2} \tag{83}$$

$$= \frac{1}{2}\left(\frac{\partial \xi}{\partial x} + \frac{\partial \zeta}{\partial x}\right); \qquad S_{23} = S_{32} = \frac{S_{yz}}{2} = \frac{1}{2}\left(\frac{\partial \zeta}{\partial y} + \frac{\partial \eta}{\partial z}\right)$$

the nine components

$$\begin{vmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{vmatrix} \tag{83}$$

will form a tensor of the second rank, as can be shown by the transformation equations of (82).

The generalized Hooke's law given by equation (22) becomes

$$T_{ij} = c_{ijk\ell} S_{k\ell} \tag{84}$$

$c_{ijk\ell}$ is a fourth rank tensor. The right hand side of the equation being the product of a fourth rank tensor by a second rank tensor is a sixth rank tensor, but since it has been contracted twice by having k and ℓ in both terms the resultant of the right hand side is a second rank tensor. Since $c_{ijk\ell}$ is a tensor of the fourth rank it will, in general, have 81 terms, but on account of the symmetry of the T_{ij} and $S_{k\ell}$ tensors, there are many equivalences between the resulting elastic constants. These equivalences can be determined by expanding the terms of (84) and comparing with the equivalent expressions of (22). For example

$$\begin{aligned} T_{11} = {} & c_{1111}S_{12} + c_{1112}S_{12} + c_{1113}S_{13} \\ & + c_{1121}S_{21} + c_{1122}S_{22} + c_{1123}S_{23} \\ & + c_{1131}S_{31} + c_{1132}S_{32} + c_{1133}S_{33}. \end{aligned} \tag{85}$$

Comparing this equation with the first of (22) noting that $S_{12} = S_{21} = \frac{S_{xy}}{2}$, etc., we have

$$\begin{aligned} & c_{1111} = c_{11}\ ;\ c_{1112} = c_{1121} = c_{16}\ ;\ c_{1133} = c_{13}\ ;\ c_{1113} = c_{1131} = c_{15}\ ; \\ & c_{1122} = c_{12}\ ;\ c_{1123} = c_{1132} = c_{14}. \end{aligned} \tag{86}$$

In a similar manner it can be shown that the elastic constants of (22) correspond to the tensor elastic constants c_{ijkl} according to the relations

$$c_{11} = c_{1111}\ ;\ c_{12} = c_{1122} = c_{2211}\ ;\ c_{13} = c_{1133} = c_{3311}\ ;\ c_{14} = c_{1123} = c_{1132} = c_{2311} = c_{3211}\ ;\ c_{15} = c_{1113} = c_{1131} = c_{1311} = c_{3111}\ ;\ c_{16} = c_{1112} = c_{1121} = c_{1211} = c_{2111}\ ;\ c_{22} = c_{2222}\ ;\ c_{23} = c_{2233} = c_{3322}\ ;\ c_{24} = c_{2223} = c_{2232} = c_{2322} = c_{3222}\ ;\ c_{25} = c_{2213} = c_{2231} = c_{1322} = c_{3122}\ ;\ c_{26} = c_{2212} = c_{2221} = c_{1222} = c_{2122}\ ;\ c_{33} = c_{3333}\ ;\ c_{34} = c_{3323} = c_{3332} = c_{2333} = c_{3233}\ ;\ c_{35} = 3313 = c_{3331} = c_{1333} = c_{3133}\ ;\ c_{36} = c_{3312} = c_{3321} = c_{1233} = c_{2133}\ ;\ c_{44} = c_{2323} = c_{2332} = c_{3223} = c_{3232}\ ;\ c_{45} = c_{2313} = c_{2331} = c_{3213} = c_{3231} = 1323 = 1332 = c_{3132} = c_{3123}\ ;\ c_{46} = c_{2312} = c_{2321} = c_{3212} = c_{3221} = c_{1223} = c_{1232} = c_{2123} = c_{2132}\ ;\ c_{55} = c_{1313} = c_{1331} = c_{3113} = c_{3131}\ ;\ c_{56} = c_{1312} = c_{1321} = c_{3112} = c_{3121} = c_{1213} = c_{1231} = c_{2113} = c_{2131}\ ;\ c_{66} = c_{1212} = c_{1221} = c_{2112} = c_{2121}\,. \tag{87}$$

Hence there are only 21 independent constants of the 81 c_{ijkl} constants which are determined from the ordinarily elastic constants c_{ij} by replacing

$$1 \text{ by } 11;\ 2 \text{ by } 22;\ 3 \text{ by } 33;\ 4 \text{ by } 23;\ 5 \text{ by } 13;\ 6 \text{ by } 12 \tag{88}$$

and taking all possible permutations of these constants by interchanging them in pairs.

The inverse elastic equations (26) can be written in the simplified form

$$S_{ij} = s_{ijkl} T_{kl}\,. \tag{89}$$

By expanding these equations and comparing with equations (26) we can establish the relationships

$$s_{11} = s_{1111}\ ;\ s_{12} = s_{1122} = s_{2211}\ ;\ s_{13} = s_{1133} = s_{3311}\ ;\ \frac{s_{14}}{2} = s_{1123} = s_{1132} = s_{2311} = s_{3211}\ ;\ \frac{s_{15}}{2} = s_{1113} = s_{1131} = s_{1311} = s_{3111}\ ;\ \frac{s_{16}}{2} = s_{1112} = s_{1121} = s_{1211} = s_{2111}\ ;\ s_{22} = s_{2222}\ ;\ s_{23} = s_{2233} = s_{3322}\ ;\ \frac{s_{24}}{2} = s_{2223} = s_{2232} = s_{2322} = s_{3222}\ ;\ \frac{s_{25}}{2} = s_{2213} = s_{2231} = s_{1322} = s_{3122}\ ;\ \frac{s_{26}}{2} = s_{2212} = s_{2221} = s_{1222} = s_{2122}\ ;\ s_{33} = s_{3333}\ ;\ \frac{s_{34}}{2} = s_{3323} = s_{3332} = s_{2333} = s_{2233}\ ;\ \frac{s_{35}}{2} = s_{3313} = s_{3331} = s_{1333} = s_{3133}\ ;\ \frac{s_{36}}{2} = \tag{90 A}$$

$$s_{3312} = s_{3321} = s_{1233} = s_{2133}\ ;\ \frac{s_{44}}{4} = s_{2323} = s_{2332} = s_{3223} = s_{3232}\ ;\ \frac{s_{45}}{4} = s_{2313} =$$

$$s_{2331} = s_{3213} = s_{3231} = s_{1323} = s_{1332} = s_{3123} = s_{3132}\ ;\ \frac{s_{46}}{4} = s_{2312} = s_{2321} = \qquad (90\ B)$$

$$s_{3212} = s_{3221} = s_{1223} = s_{1232} = s_{2123} = s_{2132}\ ;\ \frac{s_{55}}{4} = s_{1313} = s_{1331} = s_{3113} =$$

$$s_{3131}\ ;\ \frac{s_{56}}{4} = s_{1312} = s_{1321} = s_{3112} = s_{3121} = s_{1213} = s_{1231} = s_{2113} = s_{2131}\ ;$$

$$\frac{s_{66}}{4} = s_{1212} = s_{1221} = s_{2112} = s_{2121}\ .$$

Here again the $s_{ijk\ell}$ elastic constants are determined from the ordinary elastic constants s_{ij} by replacing

1 by 11, 2 by 22, 3 by 33, 4 by 23, 5 by 13, 6 by 12.

However for any number 4, 5, or 6 the elastic compliance s_{ij} has to be divided by two to equal the corresponding $s_{ijk\ell}$ compliance, and if 4, 5 or 6 occurs twice, the divisor has to be 4.

The isothermal elastic compliance of equations (39) can be expressed in tensor form

$$S_{ij} = s^{\Theta}_{ijk\ell}\, T_{k\ell} + \alpha_{ij}\, d\Theta \qquad (91)$$

where as before α_{ij} is a tensor of the second rank having the relations to the ordinary coefficients of expansion

$$\alpha_1 = \alpha_{11}\ ; \qquad \alpha_2 = \alpha_{22}\ ; \qquad \alpha_3 = \alpha_{33}\ ; \qquad \frac{\alpha_4}{2} = \alpha_{23}\ ;$$

$$\frac{\alpha_5}{2} = \alpha_{13}\ ; \qquad \frac{\alpha_6}{2} = \alpha_{12}\ .$$

The heat temperature equation of (35) is written in the simple form

$$dQ = +\alpha_{k\ell}\, T_{k\ell}\, \Theta + \rho C_p\, d\Theta. \qquad (92)$$

By eliminating $d\Theta$ from (92) and substituting in (91) the adiabatic constants are given in the simple form

$$s^{\sigma}_{ijk\ell} = s^{\Theta}_{ijk\ell} - \frac{\alpha_{ij}\alpha_{k\ell}\,\Theta}{\rho C_p}\ . \qquad (93)$$

The combination elastic and piezoelectric equations (60) can be written in the tensor form

$$S_{ij} = s^{E}_{ijk\ell} T_{k\ell} + d_{mij} E_m\ ; \qquad \delta_n = \frac{\overset{T}{\epsilon}_{mn}}{4\pi} E_m + d_{nk\ell} T_{k\ell}\ . \qquad (94)$$

Here d_{mij} is a tensor of third rank and ϵ^T_{mn} one of second rank. The d_{mij} constants are related to the eighteen ordinary constants d_{ij} by the equations

$$\begin{aligned}
&d_{11} = d_{111}\ ;\ d_{12} = d_{122}\ ;\ d_{13} = d_{133}\ ;\ \frac{d_{14}}{2} = d_{123} = d_{132}\ ;\ \frac{d_{15}}{2} = d_{113} = d_{131}\ ;\\
&\frac{d_{16}}{2} = d_{112} = d_{121}\ ;\ d_{21} = d_{211}\ ;\ d_{22} = d_{222}\ ;\ d_{23} = d_{233}\ ;\ \frac{d_{24}}{2} = d_{223} = d_{232}\ ;\\
&\frac{d_{25}}{2} = d_{213} = d_{231}\ ;\ \frac{d_{26}}{2} = d_{212} = d_{221}\ ;\ d_{31} = d_{311}\ ;\ d_{32} = d_{322}\ ;\ d_{33} = d_{333}\ ;\\
&\frac{d_{34}}{2} = d_{323} = d_{332}\ ;\ \frac{d_{35}}{2} = d_{313} = d_{331}\ ;\ \frac{d_{36}}{2} = d_{312} = d_{321}\ .
\end{aligned} \tag{95}$$

The tensor equations (94) give a simple method of expressing the piezoelectric equations in an alternate form which is useful for some purposes. This involves relating the stress, strain, and displacement, rather than the applied field strength as in (94). To do this let us multiply through the right hand equation of (94) by the tensor $4\pi\beta^T_{mn}$, obtaining

$$4\pi\beta^T_{mn}\,\delta_n = \epsilon^T_{mn}\beta^T_{mn}E_m + 4\pi d_{nk\ell}\beta^T_{mn}T_{k\ell} \tag{96}$$

where β^T_{mn} is a tensor of the "free" dielectric impermeability obtained from the determinant.

$$\beta^T_{mn} = (-1)^{(m+n)}\frac{\Delta^{\epsilon^T}_{mn}}{\Delta^{\epsilon}} \tag{97}$$

where Δ^{ϵ^T} is the determinant

$$\Delta^{\epsilon^T} = \begin{vmatrix} \epsilon^T_{11} & \epsilon^T_{12} & \epsilon^T_{13} \\ \epsilon^T_{12} & \epsilon^T_{22} & \epsilon^T_{23} \\ \epsilon^T_{13} & \epsilon^T_{23} & \epsilon^T_{33} \end{vmatrix} \tag{98}$$

and $\Delta^{\epsilon^T}_{mn}$ the minor obtained from this by suppressing the mth row and nth column. If we take the product $\epsilon^T_{mn}\beta^T_{mn}$ for the three values of m, we have as multipliers of E_1, E_2, E_3, respectively

$$\begin{aligned}
\epsilon^T_{11}\beta^T_{11} + \epsilon^T_{12}\beta^T_{12} + \epsilon^T_{13}\beta^T_{13} &= 1\\
\epsilon^T_{21}\beta^T_{21} + \epsilon^T_{22}\beta^T_{22} + \epsilon^T_{23}\beta^T_{23} &= 1\\
\epsilon^T_{31}\beta^T_{31} + \epsilon^T_{32}\beta^T_{32} + \epsilon^T_{33}\beta^T_{33} &= 1.
\end{aligned} \tag{99}$$

But by virtue of equations (97) and (98) it is obvious that the value of each term of (99) is unity. Hence we have

$$E_m = 4\pi\beta^T_{mn}\,\delta_n - (4\pi\, d_{nk\ell}\,\beta^T_{mn})\,T_{k\ell}\,. \tag{100}$$

Since the dummy index n is summed for the values 1, 2, and 3, we can set the value of the terms in brackets equal to

$$g_{mk\ell} = 4\pi\, d_{nk\ell}\, \beta^T_{mn} = 4\pi[d_{1k\ell}\, \beta^T_{m1} + d_{2k\ell}\, \beta^T_{m2} + d_{3k\ell}\, \beta^T_{m3}] \tag{101}$$

and equation (100) becomes

$$E_m = 4\pi\, \beta^T_{mn}\, \delta_n - g_{mk\ell}\, T_{k\ell}\,. \tag{102}$$

Substituting this equation in the first equations of (94) we have

$$S_{ij} = s^D_{ijk\ell}\, T_{k\ell} + g_{nij}\, \delta_n \tag{103}$$

where

$$s^D_{ijk\ell} = s^E_{ijk\ell} - d_{mij}\, g_{mk\ell} = s^E_{ijk\ell} - 4\pi[\beta^T_{mn}\, d_{nk\ell}\, d_{mij}].$$

By substituting in the various values of i, j, k and ℓ corresponding to the 21 elastic constants, the difference between the constant displacement and constant potential elastic constants can be calculated. If equations (102) and (103) are expressed in terms of the $S_1, \cdots, S_6$ strains and $T_1, \cdots, T_6$ stresses, the g_{nij} constants are related to the g_{ij} constants as are the corresponding d_{ij} constants to the d_{nij} constants of equation (95).

Another variation of the piezoelectric equations which is sometimes employed is one for which the stresses are expressed in terms of the strains and field strength. This form can be derived directly from equations (94) by multiplying both sides of the first equation by the tensor $c^E_{ijk\ell}$ for the elastic constants, where these are defined in terms of the corresponding s^E_{ij} elastic compliances by the equation

$$c^E_{ij} = (-1)^{(i+j)}\, \Delta^{sE}_{ij} / \Delta^{sE} \tag{104}$$

where Δ is the determinant

$$\Delta^{sE} = \begin{vmatrix} s^E_{11} & s^E_{12} & s^E_{13} & s^E_{14} & s^E_{15} & s^E_{16} \\ s^E_{12} & s^E_{22} & s^E_{23} & s^E_{24} & s^E_{25} & s^E_{26} \\ s^E_{13} & s^E_{23} & s^E_{33} & s^E_{34} & s^E_{35} & s^E_{36} \\ s^E_{14} & s^E_{24} & s^E_{34} & s^E_{44} & s^E_{45} & s^E_{46} \\ s^E_{15} & s^E_{25} & s^E_{35} & s^E_{45} & s^E_{55} & s^E_{56} \\ s^E_{16} & s^E_{26} & s^E_{36} & s^E_{46} & s^E_{56} & s^E_{66} \end{vmatrix}$$

and Δ^{sE}_{ij} in the minor obtained by suppressing the ith row and jth column. Carrying out the tensor multiplication we have

$$c^E_{ijk\ell}\, S_{ij} = c^E_{ijk\ell}\, s^E_{ijk\ell}\, T_{k\ell} + d_{mij}\, c^E_{ijk\ell}\, E_m\,. \tag{105}$$

As before we find that the tensor product of $c^E_{ijk\ell}\, s^E_{ijk\ell}$ is unity for all values of k and ℓ. Hence equation (105) can be written in the form

$$T_{k\ell} = c^E_{ijk\ell}\, S_{ij} - e_{mk\ell}\, E_m \tag{106}$$

where $e_{mk\ell}$ is the sum

$$e_{mk\ell} = d_{mij}\, c^E_{ijk\ell} \tag{107}$$

summed for all values of the dummy indices i and j. If we substitute the equation (106) in the last equation of (94) we find

$$\delta_n = \frac{\overset{S}{\epsilon}_{mn}}{4\pi} E_m + e_{nij} S_{ij} \tag{108}$$

where $\overset{S}{\epsilon}_{mn}$ the clamped dielectric constant is related to the free dielectric constant $\overset{T}{\epsilon}_{mn}$ by the equation

$$\overset{S}{\epsilon}_{mn} = \overset{T}{\epsilon}_{mn} - 4\pi[d_{nk\ell}\, e_{mk\ell}]. \tag{109}$$

Expressed in two index piezoelectric constants involving the strains $S_{11} \cdots S_{12}$ and stresses $T_{11} \cdots T_{12}$ the relation between the two and three index piezoelectric constants is given by the equation

$$\begin{aligned}
&e_{11} = e_{111}\ ;\ e_{12} = e_{122}\ ;\ e_{13} = e_{133}\ ;\ e_{14} = e_{123} = e_{132}\ ;\ e_{15} = e_{113} = e_{131}\ ;\\
&e_{16} = e_{112} = e_{121}\ ;\ e_{21} = e_{211}\ ;\ e_{22} = e_{222}\ ;\ e_{23} = e_{233}\ ;\ e_{24} = e_{223} = e_{232}\ ;\\
&e_{25} = e_{213} = e_{231}\ ;\ e_{26} = e_{212} = e_{221}\ ;\ e_{31} = e_{311}\ ;\ e_{32} = e_{322}\ ;\ e_{33} = e_{333}\ ;\\
&e_{34} = e_{323} = e_{332}\ ;\ e_{35} = e_{313} = e_{331}\ ;\ e_{36} = e_{312} = e_{321}\ .
\end{aligned} \tag{110}$$

Finally, the fourth form for expressing the piezoelectric relation is the one given by equation (53). Expressed in tensor form, these equations become

$$T_{k\ell} = c^D_{ijk\ell}\, S_{ij} - h_{nk\ell}\, \delta_n\ ; \quad E_m = 4\pi\beta^S_{mn}\, \delta_n - h_{mij} S_{ij} \tag{111}$$

In this equation the three index piezoelectric constants of equation (111) are related to the two index constants of equation (53) as the e constants of (110). These equations can also be derived directly from (106) and (108) by eliminating E_m from the two equations. This substitution yields the additional relations

$$\begin{aligned}
h_{nk\ell} = 4\pi e_{mk\ell}\, \beta^S_{mn}\ ; \quad c^D_{ijk\ell} &= c^E_{ijk\ell} + e_{mk\ell}\, h_{mij} = c^E_{ijk\ell}\\
&+ 4\pi\, e_{mk\ell}\, e_{nij}\, \beta^S_{mn}
\end{aligned} \tag{112}$$

where

$$\beta^S_{mn} = (-1)^{(m+n)}\, \Delta^{\epsilon^S}_{m,n} / \Delta^{\epsilon^S}$$

in which

$$\Delta^{\epsilon^S} = \begin{vmatrix} \overset{S}{\epsilon}_{11} & \overset{S}{\epsilon}_{12} & \overset{S}{\epsilon}_{13} \\ \overset{S}{\epsilon}_{12} & \overset{S}{\epsilon}_{22} & \overset{S}{\epsilon}_{23} \\ \overset{S}{\epsilon}_{13} & \overset{S}{\epsilon}_{23} & \overset{S}{\epsilon}_{33} \end{vmatrix}$$

The four forms of the piezoelectric equations, and the relation between them are given in Table I.

TABLE I

FOUR FORMS OF THE ELASTIC, DIELECTRIC, AND PIEZO ELECTRIC EQUATIONS AND THEIR INTERRELATIONS

Form	Elastic Relation	Electric Relation
1	$S_{ij} = s^E_{ijk\ell}T_{k\ell} + d_{mij}E_m$	$\delta_n = \frac{\overset{T}{\epsilon}_{mn}}{4\pi} E_n + d_{nk\ell}T_{k\ell}$
2	$S_{ij} = s^D_{ijk\ell}T_{k\ell} + g_{nij}\delta_n$	$E_m = 4\pi\beta^T_{mn}\delta_n - g_{mk\ell}T_{k\ell}$
3	$T_{k\ell} = c^E_{ij\ell k}S_{ij} - e_{mk\ell}E_m$	$\delta_n = \frac{\overset{S}{\epsilon}_{mn}}{4\pi} E_m + e_{nij}S_{ij}$
4	$T_{k\ell} = c^D_{ijk\ell}S_{ij} - h_{nk\ell}\delta_n$	$E_m = 4\pi\beta^S_{mn}\delta_n - h_{mij}S_{ij}$

Form	Relation Between Elastic Constants	Relation Between Piezoelectric Constants	Relation Between Dielectric Constants
1	$s^D_{ijk\ell} = s^E_{ijk\ell} - d_{mij}g_{mk\ell}$	$g_{mk\ell} = 4\pi\beta^T_{mn}d_{nk\ell}$	$\beta^T_{mn} = (-1)^{(m+n)}\Delta^{\epsilon^T}_{mn}/\Delta^{\epsilon^T}$
2	$c^E_{ij} = (-1)^{(i+j)}\Delta^{s^E}_{ij}/\Delta^{s^E}$	$e_{mk\ell} = d_{mij}c^E_{ijk\ell}$	$\epsilon^S_{mn} = \epsilon^T_{mn} - 4\pi(d_{nk\ell}e_{mk\ell})$
3	$c^D_{ijk\ell} = c^E_{ijk\ell} + e_{mk\ell}h_{mij}$	$h_{nk\ell} = 4\pi\beta^S_{mn}e_{mk\ell}$	$\beta^S_{mn} = \beta^T_{mn} + \frac{g_{nk\ell}h_{mk\ell}}{4\pi}$
4	$c^D_{ij} = (-1)^{(i+j)}\Delta^{s^D}_{ij}/\Delta^{s^D}$	$h_{nk\ell} = g_{nij}c^D_{ijk\ell}$	$\beta^S_{mn} = (-1)^{(m+n)}\Delta^{\epsilon^S}_{mn}/\Delta^{\epsilon^S}$

5. EFFECT OF SYMMETRY AND ORIENTATION ON THE DIELECTRIC PIEZOELECTRIC AND ELASTIC CONSTANTS OF CRYSTALS

All crystals can be divided into 32 classes depending on the type of symmetry. These groups can be divided into seven general classifications depending on how the axes are related and furthermore all 32 classes can be built out of symmetries based on twofold (binary) axes, threefold (trigonal) axes, fourfold axes of symmetry, sixfold axes of symmetry, planes of reflection symmetry and combinations of axis reflection symmetry besides a simple symmetry through the center. Each of these types of symmetry

result in a reduction of the number of dielectric, piezoelectric, and elastic constants.

Since the tensor equation is easily transformed to a new set of axes by the transformaion equations (76) this form is particularly advantageous for determining the reduction in elastic, piezoelectric and dielectric constants. For example consider the second rank tensors $\epsilon_{k\ell}$ and $\alpha_{k\ell}$ for the dielectric constant and the expansion coefficients. Ordinarily for the most general symmetry each tensor, since it is symmetrical, requires six independent coefficients. Suppose however that the X axis is an axis of twofold or binary symmetry, i.e., the properties along the positive Z axis are the same as those along the negative Z axis. If we rotate the axes 180° about the X axis so that $+Z$ is changed into $-Z$, the direction cosines are

$$\begin{aligned} \ell_1 &= \frac{\partial x_1'}{\partial x_1} = 1\,; \quad m_1 = \frac{\partial x_1'}{\partial x_2} = 0\,; \qquad n_1 = \frac{\partial x_1'}{\partial x_3} = 0 \\ \ell_2 &= \frac{\partial x_2'}{\partial x_1} = 0\,; \quad m_2 = \frac{\partial x_2'}{\partial x_2} = -1\,; \quad n_2 = \frac{\partial x_2'}{\partial x_3} = 0 \\ \ell_3 &= \frac{\partial x_3'}{\partial x_1} = 0\,; \quad m_3 = \frac{\partial x_3'}{\partial x_2} = 0\,; \qquad n_3 = \frac{\partial x_3'}{\partial x_3} = -1. \end{aligned} \tag{113}$$

The tensor transformation equations for a second rank tensor are

$$\epsilon_{ij}' = \frac{\partial x_i'}{\partial x_k}\frac{\partial x_j'}{\partial x\ell}\,\epsilon_{k\ell}\,. \tag{114}$$

Applying (113) to (114) summing for all values of k and l for each value of i, and j we have the six components

$$\epsilon_{11}' = \epsilon_{11}\,;\ \epsilon_{12}' = -\epsilon_{12}\,;\ \epsilon_{13}' = -\epsilon_{13}\,;\ \epsilon_{22}' = \epsilon_{22}\,;\ \epsilon_{23}' = \epsilon_{23}\,;\ \epsilon_{33}' = \epsilon_{33}\,. \tag{115}$$

Since a crystal having the X axis a binary axis of symmetry must have the same constants for a $+Z$ direction as for a $-Z$ direction, this condition can only be satisfied by

$$\epsilon_{12} = \epsilon_{13} = 0. \tag{116}$$

The same condition is true for the expansion coefficients since they form a second rank tensor and hence

$$\alpha_{12} = \alpha_{13} = 0. \tag{117}$$

In a third rank tensor such as d_{ijk}, e_{ijk}, g_{ijk}, h_{ijk}, we similarly find that of the eighteen independent constants

$$\begin{gathered} h_{112} = h_{16}\,;\ h_{113} = h_{15}\,;\ h_{211} = h_{21}\,;\ h_{222} = h_{22}\,;\ h_{223} = h_{24}\,; \\ h_{233} = h_{23}\,;\ h_{311} = h_{31}\,;\ h_{322} = h_{32}\,;\ h_{323} = h_{34}\,;\ h_{333} = h_{33}\,. \end{gathered} \tag{118}$$

are all zero. The same terms in the d_{ijk}, e_{ijk}, g_{ijk} tensors are also zero.

In a fourth rank tensor such as c_{ijkl}, s_{ijkl}, applying the tensor transformation equation

$$c_{ijkl} = \frac{\partial x_i'}{\partial x_m}\frac{\partial x_j'}{\partial x_n}\frac{\partial x_k'}{\partial x_o}\frac{\partial x_l'}{\partial x_p}c_{mnop} \tag{119}$$

and the condition (113) we similarly find

$$c_{15} = c_{16} = c_{25} = c_{26} = c_{35} = c_{36} = c_{45} = c_{46} = 0. \tag{120}$$

If the binary axis had been the Y axis the corresponding missing terms can be obtained by cyclically rotating the tensor indices. The missing terms are for the second, third and fourth rank tensors, transformed to two index symbols,

$$\begin{gathered}\epsilon_{23}\,,\ \epsilon_{12}\ ;\ h_{11}\,,\ h_{12}\,,\ h_{13}\,,\ h_{15}\,,\ h_{24}\,,\ h_{26}\,,\ h_{31}\,,\ h_{32}\,,\ h_{33}\,,\ h_{35}\ ;\\ c_{14}\,,\ c_{16}\,,\ c_{24}\,,\ c_{26}\,,\ c_{34}\,,\ c_{36}\,,\ c_{45}\,,\ c_{55}\,.\end{gathered} \tag{121}$$

Similarly if the Z axis is the binary axis, the missing constants are

$$\begin{gathered}\epsilon_{13}\,,\ \epsilon_{12}\ ;\ h_{11}\,,\ h_{12}\,,\ h_{13}\,,\ h_{16}\,,\ h_{21}\,,\ h_{22}\,,\ h_{23}\,,\ h_{26}\,,\ h_{34}\,,\ h_{35}\ ;\\ c_{14}\,,\ c_{15}\,,\ c_{24}\,,\ c_{25}\,,\ c_{34}\,,\ c_{35}\,,\ c_{46}\,,\ c_{56}\,.\end{gathered} \tag{122}$$

Hence a crystal of the orthorhombic bisphenoidal class or class 6, which has three binary axes, the X, Y and Z directions, will have the remaining terms,

$$\epsilon_{11}\,,\ \epsilon_{22}\,,\ \epsilon_{33}\ ;\ h_{14}\,,\ h_{25}\,,\ h_{36}\ ;\ c_{11}\,,\ c_{12}\,,\ c_{13}\,,\ c_{22}\,,\ c_{23}\,,\ c_{33}\,,\ c_{44}\,,\ c_{55}\,,\ c_{66} \tag{123}$$

with similar terms for other tensors of the same rank. Rochelle salt is a crystal of this class.

If Z is a threefold axis of symmetry, the direction cosines for a set of axes rotated 120° clockwise about Z are,

$$\begin{aligned}\ell_1 &= \frac{\partial x_1'}{\partial x_1} = -.5\,; \quad m_1 = \frac{\partial x_1'}{\partial x_2} = -.866\,; \quad n_1 = \frac{\partial x_1'}{\partial x_3} = 0\\ \ell_2 &= \frac{\partial x_2'}{\partial x_1} = .866\,; \quad m_2 = \frac{\partial x_2'}{\partial x_2} = -.5\,; \quad n_2 = \frac{\partial x_2'}{\partial x_3} = 0\\ \ell_3 &= \frac{\partial x_3'}{\partial x_1} = 0\,; \quad m_3 = \frac{\partial x_3'}{\partial x_2} = 0\,; \quad n_3 = \frac{\partial x_3'}{\partial x_3} = 1.\end{aligned} \tag{124}$$

Applying these relations to equations (114) for a second rank tensor, we find for the components

$$\begin{aligned}&\epsilon_{11}' = .25\epsilon_{11} + .433\epsilon_{12} + .75\epsilon_{22}\,;\ \epsilon_{12}' = -.433\epsilon_{11} + .25\epsilon_{12} + .433\epsilon_{22}\\ &\epsilon_{13}' = -.5\epsilon_{13} - .866\epsilon_{23}\,;\ \epsilon_{22}' = .75\epsilon_{11} - .433\epsilon_{12} + .25\epsilon_{22}\\ &\epsilon_{23}' = .866\epsilon_{13} - .5\epsilon_{23}\,;\ \epsilon_{33}' = \epsilon_{33}\,.\end{aligned} \tag{125}$$

For the third and fifth equations, since we must have $\epsilon'_{13} = \epsilon_{13}$; $\epsilon'_{23} = \epsilon_{23}$ in order to satisfy the symmetry relation, the equations can only be satisfied if

$$\epsilon_{13} = \epsilon_{23} = 0. \tag{126}$$

Similarly solving the first three equations simultaneously, we find

$$\epsilon_{12} = 0; \epsilon_{11} = \epsilon_{22}. \tag{127}$$

Hence the remaining constants are

$$\epsilon_{11} = \epsilon_{22} ; \epsilon_{33}. \tag{128}$$

Similarly for third and fourth rank tensors, for a crystal having Z a trigonal axis, the remaining terms are

$$\begin{aligned} &h_{11}, h_{12} = -h_{11}, h_{13} = 0; \ h_{14}, h_{15}, h_{16} = -h_{22} \\ &h_{21} = -h_{22}, h_{22}, h_{23} = 0, h_{24} = h_{15}; h_{25} = -h_{14}, h_{21} = -h_{11} \\ &h_{31}; h_{32} = h_{31}; h_{33}; h_{34} = 0; \ h_{35} = 0; \ h_{36} = 0 \end{aligned} \tag{129}$$

$$\begin{aligned} &c_{11} ; c_{12} ; c_{13} ; c_{14} ; c_{15} = -c_{25} ; c_{16} = 0 \\ &c_{12} ; c_{22} = c_{11} ; c_{23} = c_{13} ; c_{24} = -c_{14} ; c_{25} ; c_{26} = 0 \\ &c_{13} ; c_{25} = c_{13} ; c_{33} ; c_{34} = 0; \ c_{35} = 0; \ c_{36} = 0 \\ &c_{14} ; c_{24} = -c_{14} ; c_{34} = 0; \ c_{44} ; c_{45} = 0; \ c_{46} = c_{15} \\ &c_{15} = -c_{25} ; c_{25} ; c_{35} = 0; \ c_{45} = 0; \ c_{55} = c_{44} ; c_{56} = c_{14} \\ &c_{16} = 0; \ c_{26} = 0; \ c_{36} = 0; \ c_{46} = c_{25} ; c_{56} = c_{14} ; c_{66} = \tfrac{1}{2}(c_{11}-c_{12}). \end{aligned} \tag{130}$$

If the Z axis is a trigonal axis and the X a binary axis, as it is in quartz, the resulting constants are obtained by combining the conditions (116), (118), (120) with conditions (128), (129), (130) respectively. The resulting second, third and fourth rank tensors have the following terms

$$\begin{aligned} &\epsilon_{11} ; \epsilon_{12} = 0; \ \epsilon_{13} = 0 \\ &\epsilon_{12} = 0; \ \epsilon_{22} = \epsilon_{11} ; \epsilon_{23} = 0 \\ &\epsilon_{13} = 0; \ \epsilon_{23} = 0; \ \epsilon_{33} \end{aligned} \tag{131}$$

$$\begin{aligned} &h_{11} ; h_{12} = -h_{11} ; h_{13} = 0; \ h_{14} ; h_{15} = 0; \ h_{16} = 0 \\ &h_{21} = 0; \ h_{22} = 0; \ h_{23} = 0; \ h_{24} = 0; \ h_{25} = -h_{14} ; h_{26} = -h_{11} \\ &h_{31} = 0; \ h_{32} = 0; \ h_{33} = 0; \ h_{34} = 0; \ h_{35} = 0; \ h_{36} = 0 \end{aligned} \tag{132}$$

$$
\begin{aligned}
&c_{11}\ ;\ c_{12}\ ;\ c_{13}\ ;\ c_{14}\ ;\ c_{15}=0;\ c_{16}=0\\
&c_{12}\ ;\ c_{22}=c_{11}\ ;\ c_{23}=c_{13}\ ;\ c_{24}=-c_{14}\ ;\ c_{25}=0;\ c_{26}=0\\
&c_{13}\ ;\ c_{23}=c_{13}\ ;\ c_{33}\ ;\ c_{34}=0;\ c_{35}=0;\ c_{36}=0\\
&c_{14}\ ;\ c_{24}=-c_{14}\ ;\ c_{34}=0;\ c_{44}\ ;\ c_{45}=0;\ c_{46}=0\\
&c_{15}=0;\ c_{25}=0;\ c_{35}=0;\ c_{45}=0;\ c_{55}=c_{44}\ ;\ c_{56}=c_{14}\\
&c_{16}=0;\ c_{26}=0;\ c_{36}=0;\ c_{36}=0;\ c_{56}=c_{14}\ ;\ c_{66}=\tfrac{1}{2}\,(c_{11}-c_{12}).
\end{aligned}
\tag{133}
$$

5.1 *Second Rank Tensors for Crystal Classes*

The symmetry relations have been calculated for all classes of crystals. For a second rank tensor such as ϵ_{ij}, the following forms are required

Triclinic Classes 1 and 2

$$
\begin{vmatrix}
\epsilon_{11}\ ,\ \epsilon_{12}\ ,\ \epsilon_{13}\\
\epsilon_{12}\ ,\ \epsilon_{22}\ ,\ \epsilon_{23}\\
\epsilon_{13}\ ,\ \epsilon_{23}\ ,\ \epsilon_{33}
\end{vmatrix}
$$

Monoclinic sphenoidal, Y a binary axis, Class 3
Monoclinic domatic, Y a plane of symmetry, Class 4
Monoclinic prismatic, Center of symmetry, Class 5

$$
\begin{vmatrix}
\epsilon_{11}\ ,\ 0\ \ ,\ \epsilon_{13}\\
0\ \ ,\ \epsilon_{22}\ ,\ 0\\
\epsilon_{13}\ ,\ 0\ \ ,\ \epsilon_{33}
\end{vmatrix}
$$

Orthorhombic
Classes 6, 7, 8

$$
\begin{vmatrix}
\epsilon_{11}\ ,\ 0\ \ ,\ 0\\
0\ \ ,\ \epsilon_{22}\ ,\ 0\\
0\ \ ,\ 0\ \ ,\ \epsilon_{33}
\end{vmatrix}
\tag{134}
$$

Tetragonal, Trigonal
Hexagonal
Classes 9 to 27

$$
\begin{vmatrix}
\epsilon_{11}\ ,\ 0\ \ ,\ 0\\
0\ \ ,\ \epsilon_{11}\ ,\ 0\\
0\ \ ,\ 0\ \ ,\ \epsilon_{33}
\end{vmatrix}
$$

Cubic
Classes 28 to 32

$$
\begin{vmatrix}
\epsilon_{11}\ ,\ 0\ \ ,\ 0\\
0\ \ ,\ \epsilon_{11}\ ,\ 0\\
0\ \ ,\ 0\ \ ,\ \epsilon_{11}
\end{vmatrix}
$$

5.2 *Third Rank Tensors of the Piezoelectric Type for the Crystal Classes*

Triclinic Assymetric (Class 1) No Symmetry

$$
\begin{vmatrix}
h_{11}\ ,\ h_{12}\ ,\ h_{13}\ ,\ h_{14}\ ,\ h_{15}\ ,\ h_{16}\\
h_{21}\ ,\ h_{22}\ ,\ h_{23}\ ,\ h_{24}\ ,\ h_{25}\ ,\ h_{26}\\
h_{31}\ ,\ h_{32}\ ,\ h_{33}\ ,\ h_{34}\ ,\ h_{35}\ ,\ h_{36}
\end{vmatrix}
$$

Triclinic pinacoidal, (center of symmetry) $h = 0$ (Class 2)

Monoclinic Sphenoidal (Class 3) Y is binary axis

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & h_{16} \\ h_{21} & h_{22} & h_{23} & 0 & h_{25} & 0 \\ 0 & 0 & 0 & h_{34} & 0 & h_{36} \end{vmatrix}$$

Monoclinic domatic (Class 4) Y plane is plane of symmetry

$$\begin{vmatrix} h_{11} & h_{12} & h_{13} & 0 & h_{15} & 0 \\ 0 & 0 & 0 & h_{24} & 0 & h_{26} \\ h_{31} & h_{32} & h_{33} & 0 & h_{35} & 0 \end{vmatrix}$$

Monoclinic prismatic (center of symmetry) $h = 0$ (Class 5) (135)

Orthorhombic bisphenoidal (Class 6) X, Y, Z binary axes

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & h_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & h_{36} \end{vmatrix}$$

Orthorhombic pyramidal (Class 7) Z binary, X, Y, planes of symmetry

$$\begin{vmatrix} 0 & 0 & 0 & 0 & h_{15} & 0 \\ 0 & 0 & 0 & h_{24} & 0 & 0 \\ h_{31} & h_{32} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Orthorhombic bipyramidal (center of symmetry) $h = 0$ (Class 8)

Tetragonal bisphenoidal (Class 9) Z is quaternary alternating

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & h_{15} & 0 \\ 0 & 0 & 0 & -h_{15} & h_{14} & 0 \\ h_{31} & -h_{31} & 0 & 0 & 0 & h_{36} \end{vmatrix}$$

Tetragonal pyramidal (Class 10) Z is quaternary

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & h_{15} & 0 \\ 0 & 0 & 0 & h_{15} & -h_{14} & 0 \\ h_{31} & h_{31} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Tetragonal scalenohedral (Class 11) Z quaternary, X and Y binary

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & h_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & h_{36} \end{vmatrix}$$

Tetragonal trapezohedral (Class 12) Z quaternary, X and Y binary

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -h_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Tetragonal bipyramidal (center of symmtery) $h = 0$ (Class 13)

Ditetragonal pyramidal (Class 14) Z quaternary, X and Y planes of symmetry

$$\begin{vmatrix} 0 & 0 & 0 & 0 & h_{15} & 0 \\ 0 & 0 & 0 & h_{15} & 0 & 0 \\ h_{31} & h_{31} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Ditetragonal bipyramidal (center of symmetry) $h = 0$ (Class 15)

Trigonal pyramidal (Class 16) Z trigonal axis

$$\begin{vmatrix} h_{11} & -h_{11} & 0 & h_{14} & h_{15} & -h_{22} \\ -h_{22} & h_{22} & 0 & h_{15} & -h_{14} & -h_{11} \\ h_{31} & h_{31} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Trigonal rhombohedral (Class 17) center of symmetry, $h = 0$

Trigonal trapezohedral (Class 18), Z trigonal, X binary

$$\begin{vmatrix} h_{11} & -h_{11} & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -h_{14} & -h_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Trigonal bipyramidal (Class 19), Z trigonal, plane of symmetry

$$\begin{vmatrix} h_{11} & -h_{11} & 0 & 0 & 0 & -h_{22} \\ -h_{22} & h_{22} & 0 & 0 & 0 & -h_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Ditrigonal pyramidal (Class 20) Z trigonal, Y plane of symmetry

$$\begin{vmatrix} 0 & 0 & 0 & 0 & h_{15} & -h_{22} \\ -h_{22} & h_{22} & 0 & h_{15} & 0 & 0 \\ h_{31} & h_{31} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Ditrigonal scalenohedral (Class 21) center of symmetry, $h = 0$

Ditrigonal bipyramidal (Class 22) Z trigonal, Z plane of symmetry and Y plane of symmetry

$$\begin{vmatrix} h_{11} & -h_{11} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -h_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal pyramidal (Class 23) Z hexagonal

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & h_{15} & 0 \\ 0 & 0 & 0 & h_{15} & -h_{14} & 0 \\ h_{31} & h_{31} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal trapezohedral (Class 24) Z hexagonal, X binary

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -h_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal bipyramidal (Class 25) center of symmetry, $h = 0$

Dihexagonal pyramidal (Class 26) X hexagonal Y plane of symmetry

$$\begin{vmatrix} 0 & 0 & 0 & 0 & h_{15} & 0 \\ 0 & 0 & 0 & h_{15} & 0 & 0 \\ h_{31} & h_{31} & h_{33} & 0 & 0 & 0 \end{vmatrix}$$

Dihexagonal bipyramidal (Class 27) center of symmetry, $h = 0$

Cubic tetrahedral-pentagonal-dedocahedral (Class 28) X, Y, Z binary

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & h_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & h_{14} \end{vmatrix}$$

Cubic pentagonal-icositetetrahedral (Class 29) $h = 0$

Cubic, dyakisdodecahedral (Class 30) center of symmetry, $h = 0$

Cubic, hexakistetrahedral (Class 31) X, Y, Z quaternary alternating

$$\begin{vmatrix} 0 & 0 & 0 & h_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & h_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & h_{14} \end{vmatrix}$$

Cubic, hexakis-octahedral (Class 32) center of symmetry, $h = 0$

This third rank tensor has been expressed in terms of two index symbols rather than the three index tensor symbols, since the two index symbols are commonly used in expressing the piezoelectric effect. The relations for the h and e constants are

$$h_{i4}, h_{i5}, h_{i6} \text{ are equivalent to } h_{i23}, h_{i13}, h_{i12} \tag{136}$$

in three index symbols, whereas for the d_{ij} and g_{ij} constants we have the relations

$$\frac{d_{i4}}{2}, \quad \frac{a_{i5}}{2}, \quad \frac{d_{i6}}{2} \text{ are equivalent to } d_{i23}, d_{i13}, d_{i12} \tag{137}$$

Hence the d_i relations for classes 16, 18, 19, and 22 will be somewhat different than the h symbols given above. These classes will be

Class 16

$$\begin{vmatrix} d_{11} & -d_{11} & 0 & d_{14} & d_{15} & -2d_{22} \\ -d_{22} & d_{22} & 0 & d_{15} & -d_{14} & -2d_{11} \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{vmatrix}$$

Class 18

$$\begin{vmatrix} d_{11} & -d_{11} & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -d_{14} & -2d_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

(138)

Class 19

$$\begin{vmatrix} d_{11} & -d_{11} & 0 & 0 & 0 & -2d_{22} \\ -d_{22} & d_{22} & 0 & 0 & 0 & -2d_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Class 22

$$\begin{vmatrix} d_{11} & -d_{11} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2d_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

5.3 *Fourth Rank Tensors of the Elastic Type for the Crystal Classes*

Triclinic System (Classes 1 and 2) 21 moduli

$$\begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{12} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{13} & c_{23} & c_{33} & c_{34} & c_{35} & c_{36} \\ c_{14} & c_{24} & c_{34} & c_{44} & c_{45} & c_{46} \\ c_{15} & c_{25} & c_{35} & c_{45} & c_{55} & c_{56} \\ c_{16} & c_{26} & c_{36} & c_{46} & c_{56} & c_{66} \end{vmatrix}$$

The s tensor is entirely analogous

(139)

Monoclinic System (Classes 3, 4 and 5) 12 moduli

$$\begin{vmatrix} c_{11} & c_{12} & c_{13} & 0 & c_{15} & 0 \\ c_{12} & c_{22} & c_{23} & 0 & c_{25} & 0 \\ c_{13} & c_{23} & c_{33} & 0 & c_{35} & 0 \\ 0 & 0 & 0 & c_{44} & 0 & c_{46} \\ c_{15} & c_{25} & c_{35} & 0 & c_{55} & 0 \\ 0 & 0 & 0 & c_{46} & 0 & c_{66} \end{vmatrix}$$

The s tensor is entirely analogous

Rhombic System (Classes 6, 7 and 8) 9 moduli	$\begin{matrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{22} & c_{23} & 0 & 0 & 0 \\ c_{13} & c_{23} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{matrix}$	The s tensor is entirely analogous
Tetragonal system, Z a fourfold axis (Classes 9, 10, 13) 7 moduli	$\begin{matrix} c_{11} & c_{12} & c_{13} & 0 & 0 & c_{16} \\ c_{12} & c_{11} & c_{13} & 0 & 0 & -c_{16} \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ c_{16} & -c_{16} & 0 & 0 & 0 & c_{66} \end{matrix}$	The s tensor is entirely analogous
Tetragonal system, Z a fourfold axis, X a twofold axis (Classes 11, 12, 14, 15) 6 moduli	$\begin{matrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{matrix}$	The s tensor is entirely analogous
Trigonal system, Z a twofold axis, (Classes 16, 17) 7 moduli	$\begin{matrix} c_{11} & c_{12} & c_{13} & c_{14} & -c_{25} & 0 \\ c_{12} & c_{11} & c_{13} & -c_{14} & c_{25} & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ c_{14} & -c_{14} & 0 & -c_{44} & 0 & c_{25} \\ -c_{25} & c_{25} & 0 & 0 & c_{44} & c_{14} \\ 0 & 0 & 0 & c_{25} & c_{14} & \frac{c_{11}-c_{12}}{2} \end{matrix}$	The s tensor is analogous except that $s_{46} = 2s_{25}$, $s_{56} = 2s_{14}$, $s_{66} = 2(s_{11} - s_{12})$

Trigonal system, Z a trigonal axis, X a binary axis (Classes 18, 20, 21) 6 moduli	$\begin{matrix} c_{11} & c_{12} & c_{13} & c_{14} & 0 & 0 \\ c_{12} & c_{11} & c_{13} & -c_{14} & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ c_{14} & -c_{14} & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & c_{14} \\ 0 & 0 & 0 & 0 & c_{14} & \frac{c_{11}-c_{12}}{2} \end{matrix}$	The s tensor is analogous except that $s_{56} = 2s_{14}$, $s_{66} = 2(s_{11} - s_{12})$
Hexagonal system, Z a sixfold axis, X a twofold axis (Classes 19, 22, 23, 24, 25, 26, 27) 5 moduli	$\begin{matrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{c_{11}-c_{12}}{2} \end{matrix}$	The s tensor is analogous except $s_{66} = 2(s_{11} - s_{12})$
Cubic system (Classes 28, 29, 30, 31, 32) 3 moduli	$\begin{matrix} c_{11} & c_{12} & c_{12} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{12} & 0 & 0 & 0 \\ c_{12} & c_{12} & c_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{44} \end{matrix}$	The s tensor is entirely analogous
Isotropic bodies, 2 moduli	$\begin{matrix} c_{11} & c_{12} & c_{12} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{12} & 0 & 0 & 0 \\ c_{12} & c_{12} & c_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{c_{11}-c_{12}}{2} & & \\ 0 & 0 & 0 & 0 & \frac{c_{11}-c_{12}}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{c_{11}-c_{12}}{2} \end{matrix}$	The s tensor analogous except last three diagonal terms are $2(s_{11} - s_{12})$

5.4 *Piezoelectric Equations for Rotated Axes*

Another application of the tensor equations for rotated axes is in determining the piezoelectric equations of crystals whose length, width, and thickness do not coincide with the crystallographic axes of the crystal. Such oriented cuts are useful for they sometimes give properties that cannot be obtained with crystals lying along the crystallographic axes. Such properties may be higher electromechanical coupling, freedom from coupling to undesired modes of motion, or low temperature coefficients of frequency. Hence in order to obtain the performance of such crystals it is necessary to be able to express the piezoelectric equations in a form suitable for these orientations. In fact in first measuring the properties of these crystals a series of oriented cuts is commonly used since by employing such cuts the resulting frequencies, and impedances can be used to calculate all the primary constants of the crystal.

The piezoelectric equations (111) are

$$T_{k\ell} = c^D_{ijk\ell} S_{ij} - h_{nk\ell}\delta_n \; ; \quad E_m = 4\pi\beta^S_{mn}\delta_n - h_{mij}S_{ij} \; . \tag{111}$$

The first equation is a tensor of the second rank, while the second equation is a tensor of the first rank. If we wish to transform these equations to another set of axes x', y', z', we can employ the tensor transformation equations

$$T'_{k\ell} = \frac{\partial x'_k}{\partial x_k}\frac{\partial x'_\ell}{\partial x_\ell} T_{k\ell} = \frac{\partial x'_k}{\partial x_k}\frac{\partial x'_\ell}{\partial x_\ell}$$

$$[c^D_{11k\ell} S_{11} + 2c^D_{12k\ell} S_{12} + 2c^D_{13k\ell} S_{15} + c^D_{22k\ell} S_{22}$$

$$+ 2c^D_{23k\ell} S_{23} + c^D_{33k\ell} S_{33}] - \frac{\partial x'_k}{\partial x_k}\frac{\partial x'_\ell}{\partial x_\ell}[h_{1k\ell}\delta_1 + h_{2k\ell}\delta_2 + h_{3k\ell}\delta_3] \tag{140}$$

$$E'_m = 4\pi \frac{\partial x'_m}{\partial x_m}[\beta^S_{m1}\delta_1 + \beta^S_{m2}\delta_2 + \beta^S_{m3}\delta_3] - \frac{\partial x'_m}{\partial x_m}$$

$$[h_{m11}S_{11} + 2h_{m12}S_{12} + 2h_{m13}S_{13} + h_{m22}S_{22} + 2h_{m23}S_{23} + h_{m33}S_{33}].$$

These equations express the new stresses and fields in terms of the old strains and displacements. To complete the transformation we need to express all quantities in terms of the new axes. For this purpose we employ the tensor equations

$$S_{ij} = \frac{\partial x_i}{\partial x'_i}\frac{\partial x_j}{\partial x'_j} S'_{ij} \; ; \qquad \delta_n = \frac{\partial x_n}{\partial x'_n}\delta'_n \tag{141}$$

where $\frac{\partial x_i}{\partial x'_i}$ are the direction cosines between the old and new axes. It is obvious that $\frac{\partial x'_i}{\partial x_i} = \frac{\partial x_i}{\partial x'_i}$ and the relations can be written

$$\begin{aligned}
\ell_1 &= \frac{\partial x_1'}{\partial x_1} = \frac{\partial x_1}{\partial x_1'}; \qquad \ell_2 = \frac{\partial x_1}{\partial x_2'}; \qquad \ell_3 = \frac{\partial x_1}{\partial x_3'} \\
m_1 &= \frac{\partial x_2}{\partial x_1'}; \qquad m_2 = \frac{\partial x_2}{\partial x_2'}; \qquad m_3 = \frac{\partial x_2}{\partial x_3'} \\
n_1 &= \frac{\partial x_3}{\partial x_1'}; \qquad n_2 = \frac{\partial x_3}{\partial x_2'}; \qquad n_3 = \frac{\partial x_3}{\partial x_3'}.
\end{aligned} \tag{142}$$

Hence substituting equations (141) in equations (140) the transformation equations between the new and old axes become

$$\begin{aligned}
T_{k\ell}' &= c_{ijk\ell}^{D} \frac{\partial x_k'}{\partial x_k} \frac{\partial x_\ell'}{\partial x_\ell} \frac{\partial x_i}{\partial x_i'} \frac{\partial x_j}{\partial x_j'} S_{ij}' - h_{mk\ell} \frac{\partial x_k'}{\partial x_k} \frac{\partial x_\ell'}{\partial x_\ell} \frac{\partial x_n}{\partial x_n'} \delta_n' \\
E_m' &= 4\pi\beta_{mn}^{S} \frac{\partial x_m'}{\partial x_m} \frac{\partial x_n}{\partial x_n'} \delta_n' - h_{mij} \frac{\partial x_m'}{\partial x_m} \frac{\partial x_i}{\partial x_i'} \frac{\partial x_j}{\partial x_j'} S_{ij}'.
\end{aligned} \tag{143}$$

These equations then provide means for determining the transformation of constants from one set of axes to another.

As an example let us consider the case of an ADP crystal, vibrating longitudinally with its length along the x_1' axis, its width along the x_2' axis and its thickness along the x_3' axis, which is also the x_3 axis, and determine the elastic, piezoelectric and dielectric constants that apply for this cut when x_1' is $\Theta = 45°$ from x_1. Under these conditions

$$\begin{aligned}
\ell_1 &= \frac{\partial x_1'}{\partial x_1} = \frac{\partial x_1}{\partial x_1'} = \cos\theta; \qquad \ell_2 = \frac{\partial x_2'}{\partial x_1} = \frac{\partial x_1}{\partial x_2'} = -\sin\theta; \\
&\qquad\qquad \ell_3 = \frac{\partial x_3'}{\partial x_1} = \frac{\partial x_1}{\partial x_3'} = 0 \\
m_1 &= \frac{\partial x_1'}{\partial x_2} = \frac{\partial x_2}{\partial x_1'} = \sin\theta; \qquad m_2 = \frac{\partial x_2'}{\partial x_2} = \frac{\partial x_2}{\partial x_2'} = \cos\theta; \\
&\qquad\qquad m_3 = \frac{\partial x_3'}{\partial x_2} = \frac{\partial x_2}{\partial x_3'} = 0 \\
n_1 &= \frac{\partial x_1'}{\partial x_3} = \frac{\partial x_3}{\partial x_1'} = 0; \qquad n_2 = \frac{\partial x_2'}{\partial x_3} = \frac{\partial x_3}{\partial x_2'} = 0; \\
&\qquad\qquad n_3 = \frac{\partial x_3'}{\partial x_3} = \frac{\partial x_3}{\partial x_3'} = 1.
\end{aligned} \tag{144}$$

Since ADP belongs to the orthorhombic bisphenoidal (Class 6), it will have the dielectric, piezoelectric and elastic tensors shown by equations (134), (135), (139). Applying equations (143) and (144) to these tensors it is

readily shown that the stresses for $\Theta = 45^\circ$ are given by the equations expressed in two index symbols

$$T_1' = \frac{(c_{11}^D + c_{12}^D + 2c_{66}^D)}{2} S_1' + \frac{(c_{11}^D + c_{12}^D - 2c_{66}^D)}{2} S_2' + c_{13}^D S_3' - h_{36}\delta_3'$$

$$T_2' = \frac{(c_{11}^D + c_{12}^D - 2c_{66}^D)}{2} S_1' + \frac{(c_{11}^D + c_{12}^D + 2c_{66}^D)}{2} S_2' + c_{13}^D S_3' + h_{36}\delta_3' \qquad (145)$$

$$T_3' = c_{13}^D S_1' + c_{13}^D S_2' + c_{33} S_3'$$

$$T_4' = c_{44}^D S_4' + h_{14}\delta_2' ; \qquad E_1 = -h_{14}S_5' + 4\pi[\beta_{11}\delta_1']$$

$$T_5' = c_{44}^D S_5' - h_{14}\delta_1' ; \qquad E_2 = h_{14}S_1' + 4\pi[\beta_{11}\delta_2']$$

$$T_6' = \frac{(c_{11}^D - c_{12}^D)}{2} S_6' ; \qquad E_3 = -h_{36}[S_1' - S_2'] + 4\pi[\beta_{33}\delta_3'].$$

For a long thin longitudinally vibrating crystal all the stresses are zero except the stress T_1' along the length of the crystal. Hence it is more advantageous to use equations which express the strains in terms of the stresses since all the stresses can be set equal to zero except T_1'. All the strains are then dependent functions of the strain S_1' and this only has to be solved for. Furthermore, since plated crystals are usually used to determine the properties of crystals, and the field perpendicular to a plated surface is zero, the only field existing in a thin crystal will be E_3' if the thickness is taken along the x_3' or Z' axis. Hence the equations that express the strains in terms of the stresses and fields are more advantageous for calculating the properties of longitudinally vibrating crystals. By orienting such crystals with respect to the crystallographic axis, all of the elastic constants except the shear elastic constants can be determined. All of the piezoelectric and dielectric constants can be determined from measurements on oriented longitudinally vibrating crystals.

For such measurements it is necessary to determine the appropriate elastic, piezoelectric, and dielectric constants for a crystal oriented in any direction with respect to the crystallographic axes. We assume that the length lies along the x_1' axis, the width along the x_2' axis and the thickness along the x_3' axis. Starting with equations of the form

$$S_{ij} = s_{ijk\ell}^E T_{k\ell} + d_{ijm}E_m$$
$$\delta_n = \frac{\epsilon_{mn}^T}{4\pi} E_m + d_{nk\ell} T_{k\ell} \qquad (146)$$

and transforming to a rotated system of axes whose direction cosines are given by (142), the resulting equation becomes

$$S'_{ij} = s^E_{ijk\ell}\frac{\partial x'_i}{\partial x_i}\frac{\partial x'_j}{\partial x_j}\frac{\partial x_k}{\partial x'_k}\frac{\partial x_\ell}{\partial x'_\ell}T'_{k\ell} + d_{ijm}\frac{\partial x'_i}{\partial x_i}\frac{\partial x'_j}{\partial x_j}\frac{\partial x_m}{\partial x'_m}E'_m,$$
$$\delta'_n = \frac{\epsilon^T_{mn}}{4\pi}\frac{\partial x'_n}{\partial x_n}\frac{\partial x_m}{\partial x'_m}E'_m + d_{nk\ell}\frac{\partial x'_n}{\partial x_n}\frac{\partial x_k}{\partial x'_k}\frac{\partial x_\ell}{\partial x'_\ell}T'_{k\ell}. \tag{147}$$

All the stresses except T'_{11} can be set equal to zero and all the fields except E'_3 vanish. Furthermore, all the strains are dependently related to S'_{11}. Hence for a thin longitudinal crystal the equation of motion becomes

$$S'_{11} = s^E_{ijk\ell}\frac{\partial x'_1}{\partial x_i}\frac{\partial x'_1}{\partial x_j}\frac{\partial x_k}{\partial x'_1}\frac{\partial x_\ell}{\partial x'_1}T'_{11} + d_{ijm}\frac{\partial x'_1}{\partial x_i}\frac{\partial x'_1}{\partial x_j}\frac{\partial x_m}{\partial x'_3}E'_m,$$
$$\delta'_3 = \frac{\epsilon^T_{mn}}{4\pi}\frac{\partial x'_3}{\partial x_n}\frac{\partial x_m}{\partial x'_3}E'_3 + d_{nk\ell}\frac{\partial x'_3}{\partial x_n}\frac{\partial x_k}{\partial x'_1}\frac{\partial x_\ell}{\partial x'_1}T'_{11}. \tag{148}$$

In terms of the two index symbols for the most general type of crystal, we have

$$\begin{aligned}
s^{E'}_{1111} = s^{E'}_{11} &= s^E_{11}\ell_1^4 + (2s^E_{12} + s^E_{66})\ell_1^2m_1^2 + (2s^E_{13} + s^E_{55})\ell_1^2n_1^2 \\
&+ 2(s^E_{14} + s^E_{56})\ell_1^2m_1n_1 + 2s^E_{15}\ell_1^3n_1 + 2s^E_{16}\ell_1^3m_1 + s^E_{22}m_1^4 \\
&+ (2s^E_{23} + s^E_{44})m_1^2n_1^2 + 2s^E_{24}m_1^3n_1 + 2(s^E_{25} + s^E_{46})m_1^2\ell_1n_1 \\
&+ 2s^E_{26}m_1^3\ell_1 + s^E_{33}n_1^4 + 2s^E_{34}n_1^3m_1 + 2s^E_{35}n_1^3\ell_1 \\
&+ 2(s^E_{36} + s^E_{45})n_1^2\ell_1m_1 \\
d'_{111} = d'_{11} &= d_{11}\ell_3\ell_1^2 + d_{12}\ell_3m_1^2 + d_{13}\ell_3n_1^2 + d_{14}\ell_3m_1n_1 + d_{15}\ell_3\ell_1n_1 \\
&+ d_{16}\ell_3\ell_1m_1 + d_{21}m_3\ell_1^2 + d_{22}m_3m_1^2 + d_{23}m_3n_1^2 + d_{24}m_3m_1n_1 \\
&+ d_{25}n_3\ell_1n_1 + d_{26}m_3\ell_1m_1 + d_{31}n_3\ell_1^2 + d_{32}n_3m_1^2 + d_{33}n_3n_1^2 \\
&+ d_{34}n_3m_1n_1 + d_{35}n_3\ell_1n_1 + d_{36}n_3\ell_1m_1 \\
\epsilon^{T'}_{33} &= \epsilon^T_{11}\ell_3^2 + 2\epsilon^T_{12}\ell_3m_3 + 2\epsilon^T_{13}\ell_3n_3 + \epsilon^T_{22}m_3^2 + 2\epsilon^T_{23}m_3n_3 + \epsilon^T_{33}n_3^2
\end{aligned} \tag{149}$$

Hence by cutting 18 crystals with independent direction cosines 9 elastic constants and 6 relations between the remaining twelve constants can be determined. All of the piezoelectric constants and all of the dielectric constants can be determined from these measurements. These constants can be measured by measuring the resonant and antiresonant frequencies and the capacity at low frequencies. The resonant frequency f_R is determined by the formula

$$f_R = \frac{1}{2\ell}\sqrt{\frac{1}{\rho s^{E'}_{11}}} \tag{150}$$

for any long thin crystal vibrating longitudinally. Hence when the density is known, $s_{11}^{E'}$ can be calculated from the resonant frequency and the length of the crystal. Using the values of $s_{11}^{E'}$ obtained for 15 independent orientations enough data is available to solve for the constants of the first of equations (149). The capacities of the different crystal orientations measured at low frequencies determine the dielectric constant $\epsilon_{33}^{T'}$ and six orientations are sufficient to determine the six independent dielectric constants ϵ_{mn}^{T}. The separation between resonance and antiresonance $\Delta f = f_A - f_R$ determines the piezoelectric constant d_{11}' according to the formula

$$d_{11}' \doteq \frac{\pi}{2}\sqrt{\frac{\Delta f}{f_R}}\sqrt{\frac{\epsilon_{33}^{T'}}{4\pi}\, s_{11}^{E'}}\,. \tag{151}$$

The values of d_{11}' measured for 18 independent orientations are sufficient to determine the eighteen independent piezoelectric constants.

The remaining six elastic constants can be determined by measuring long thin crystals in a face shear mode of motion. Since this is a contour mode of motion, the equations are considerably more complicated than for a longitudinal mode and involve elastic constants that are not constant field or constant displacement constants. It can be shown[5] that the fundamental frequency of a crystal with its length along x_1, width (frequency determining direction) along x_2 and thickness (direction of applied field) along x_3, will be

$$f = \frac{1}{2\ell_w}\sqrt{\frac{c_{22}^{c,E} + c_{66}^{c,E} \pm \sqrt{(c_{22}^{c,E} - c_{66}^{c,E})^2 + 4c_{26}^{c,E^2}}}{2\rho}} \tag{152}$$

where the contour elastic constants are given in terms of the fundamental elastic constants by

$$\begin{aligned} c_{22}^{c,E} &= \frac{s_{11}^{E} s_{66}^{E} - s_{16}^{E^2}}{\Delta}\,; \qquad c_{26}^{c,E} = \frac{s_{12}^{E} s_{16}^{E} - s_{11}^{E} s_{26}^{E}}{\Delta}\,; \\ c_{66}^{c,E} &= \frac{s_{11}^{E} s_{22}^{E} - s_{12}^{E^2}}{\Delta} \end{aligned} \tag{153}$$

where Δ is the determinant

$$\Delta = \begin{vmatrix} s_{11}^{E}, & s_{12}^{E}, & s_{16}^{E} \\ s_{12}^{E}, & s_{22}^{E}, & s_{26}^{E} \\ s_{16}^{E}, & s_{26}^{E}, & s_{66}^{E} \end{vmatrix} \tag{154}$$

Since all of the constants except s_{12}^{E} and s_{66}^{E} can be determined by measurements on longitudinal crystals and the value of $(2s_{12}^{E} + s_{66}^{E})$ has been de-

[5] This is proved in a recent paper "Properties of Dipotassium Tartrate (DKT) Crystals," *Phys. Rev., Nov., 1946.*

termined, the measurement of the lowest mode of the face shear crystal gives one more relation and hence the values of s_{12}^E and s_{66}^E can be uniquely determined.

Similar measurements with crystals cut normal to x_1 and width along x_3 and with crystals cut normal to x_2 and width along x_1 determine the constants s_{44}^E, s_{23}^E and s_{55}^E, s_{13}^E respectively. The equivalent constants are obtained by adding 1 to each subscript 1, 2, 3 or 4, 5, 6 for the first crystal with the understanding that $3 + 1 = 1$ and $6 + 1 = 4$. For the second crystal 2 is added to each subscript.

Finally the remaining three constants can be determined by measuring the face shear mode of three crystals that have their lengths along one of the crystallographic axes and their width (frequency determining axis) 45° from the other two axes.

Any symmetry existing in the crystal will cut down on the number of constants and hence on the number of orientations to determine the fundamental constants.

6. Temperature Effects in Crystals

In section 2 a general expression was developed for the effects of temperature and entropy on the constants of a crystal. Two methods were given, one which considers the stresses, field, and temperature differentials as the independent variables, and the second which considers the strains, displacements and entropy as the independent variables. In tensor form the 10 equations for the first method take the form

$$\begin{aligned} T_{k\ell} &= c_{ijk\ell}^{D,\sigma} S_{ij} - h_{nk\ell}^{\sigma} \delta_n - \lambda_{k\ell}^{S,D}\, dQ \\ E_m &= -h_{mij}^{\sigma} S_{ij} + 4\pi\beta_{mn}^{S,\sigma} \delta_n - q_m^{S,D}\, dQ \\ d\Theta &= -\Theta\lambda_{ij}^{S,D} S_{ij} - \Theta q_n^{S,D} \delta_n + \frac{dQ}{\rho c_v^{D}} \end{aligned} \tag{155}$$

The piezoelectric relations have already been discussed for adiabatic conditions assuming that no increments of heat $d'Q$ have been added to the crystal.

If now an increment of heat dQ is suddenly added to any element of the crystal, the first equation shows that a sudden expansive stress is generated proportional to the constant $\lambda_{k\ell}^{S,D}$ which has to be balanced by a negative stress (a compression) in order that no strain or electric displacement shall be generated. This effect can be called the stress caloric effect. The second equation of (155) shows that if an increment of heat dQ is added to the crystal an inverse field E_m has to be added if the strain and surface charge are to remain unchanged. This effect may be called the field caloric

effect. Finally the third equation of (155) shows that there is a reciprocal effect in which a stress or a displacement generates a change in temperature even in the absence of added heat dQ. These effects can be called the strain temperature and charge temperature effects.

The second form of the piezoelectric equations given by (58) are more familiar. In tensor form these can be written

$$\begin{aligned} S_{ij} &= s^{E,\Theta}_{ijk\ell} T_{k\ell} + d^{\Theta}_{mij} E_m + \alpha^{E}_{ij}\, d\Theta \\ \delta_n &= d^{\Theta}_{nk\ell} T_{k\ell} + \frac{\epsilon^{T,\Theta}_{mn}}{4\pi} E_m + p^{T}_{n}\, d\Theta \\ dQ = \Theta\, d\sigma &= \Theta \alpha^{E}_{k\ell} T_{k\ell} + \Theta p^{T}_{m} E_m + \rho C^{E}_{p}\, d\Theta \end{aligned} \tag{156}$$

The α^{E}_{ij} are the temperature expansion coefficients measured at constant field. In general these are a tensor of the second rank having six components. The constants p^{T}_{n} are the pyroelectric constants measured at displacements which relate the increase in polarization or surface charge due to an increase in temperature. They are equal to the so-called "true" pyroelectric constants which are the polarizations at constant volume caused by an increase in temperature plus the "false" pyroelectric effect of the first kind which represents the polarization caused by a uniform temperature expansion of the crystal as its temperature increases by $d\Theta$. As mentioned previously it is more logical to call the two effects the pyroelectric effects at constant stress and constant strain. By eliminating the stresses from the first of equations (156) and substituting in the second equation it is readily shown that

$$p^{S}_{n} = p^{T}_{n} - \alpha^{E}_{ij} e^{\Theta}_{nij} \tag{157}$$

Hence the difference between the pyroelectric effect at constant stress and the pyroelectric effect at constant strain is the so-called "false" pyroelectric effect of the first kind $\alpha^{E}_{ij} e^{\Theta}_{nij}$.

The first term on the right side of the last equation is called the heat of deformation, for it represents the heat generated by the application of the stresses $T_{k\ell}$. The second term is called the electrocaloric effect and it represents the heat generated by the application of a field. The last term is p times the specific heat at constant pressure and constant field.

The temperature expansion coefficients α^{E}_{ij} form a tensor of the second rank and hence have the same components for the various crystal classes as do the dielectric constants shown by equation (134).

The pyroelectric tensor p^{T}_{n} and p^{S}_{n} are tensors of the first rank and in general will have three components p_1, p_2, and p_3. In a similar manner to that used for second, third and fourth rank tensors it can be shown that the various crystal classes have the following components for the first rank tensor p_n.

Class 1: components p_1, p_2, p_3.

Class 3: Y axis of binary symmetry, components 0, p_2, 0 (158)

Class 4: components p_1, 0, p_3.

Classes 7, 10, 14, 16, 20, 23, and 26: components 0, 0, p_3; and Classes 2, 5, 6, 8, 9, 11, 12, 13, 15, 17, 18, 19, 21, 22, 24, 25, 27, 28, 29, 30, 31, and 32: components 0, 0, 0, i.e., $p = 0$.

For a hydrostatic pressure, the stress tensor has the components

$$T_{11} = T_{22} = T_{33} = -p = \text{pressure}; \qquad T_{12} = T_{13} = T_{23} = 0 \quad (159)$$

Hence the displacement equations of (156) can be written in the form

$$\delta_n = \frac{\epsilon^{T,\Theta}_{mn}}{4\pi} E_m - d^{\Theta}_n p + p^T_n d\Theta \quad (160)$$

where

$$d^{\Theta}_n p = d^{\Theta}_{n11} T_{11} + d^{\Theta}_{n22} T_{22} + d^{\Theta}_{n33} T_{33}$$

that is the contracted tensor $d_{nkk} T_{kk}$. This is a tensor of the first rank which has the same components as the pyroelectric tensor p_n for the various crystal classes.

7. Second Order Effects in Piezoelectric Crystals

We have so far considered only the conditions for which the stresses and fields are linear functions of the strains and electric displacements. A number of second order effects exist when we consider that the relations are not linear. Such relations are of some interest in ferroelectric crystals such as Rochelle salt. A ferroelectric crystal is one in which a spontaneous polarization exists over certain temperature ranges due to a cooperative effect in the crystal which lines up all of the elementary dipoles in a given "domain" all in one direction. Since a spontaneous polarization occurs in such crystals it is more advantageous to develop the equations in terms of the electric displacement rather than the external field. Also heat effects are not prominent in second order effects so that we develop the strains and potentials in terms of the stresses and electric displacements D. By means of McLaurin's theorem the first and second order terms are in tensor form

$$\begin{aligned} S_{ij} &= \frac{\partial S_{ij}}{\partial T_{k\ell}} T_{k\ell} + \frac{\partial S_{ij}}{\partial \delta_n} \delta_n + \frac{1}{2!}\left[\frac{\partial^2 S_{ij}}{\partial T_{k\ell}\partial T_{qr}} T_{k\ell} T_{qr}\right. \\ &\quad \left. + 2\frac{\partial^2 S_{ij}}{\partial T_{k\ell}\partial \delta_n} T_{k\ell}\delta_n + \frac{\partial^2 S_{ij}}{\partial \delta_n \partial \delta_0}\delta_n \delta_0\right] + \cdots \text{higher terms} \\ E_m &= \frac{\partial E_m}{\partial T_{k\ell}} T_{k\ell} + \frac{\partial E_m}{\partial \delta_n} \delta_n + \frac{1}{2!}\left[\frac{\partial^2 E_m}{\partial T_{k\ell}\partial T_{qr}} T_{k\ell} T_{qr}\right. \\ &\quad \left. + 2\frac{\partial^2 E_m}{\partial T_{k\ell}\partial \delta_n} T_{k\ell}\delta_n + \frac{\partial^2 E_m}{\partial \delta_n \partial \delta_0}\delta_n \delta_0\right] + \cdots \text{higher terms} \end{aligned} \quad (161)$$

whereas before $\delta = D/4\pi$.

In this equation the linear partial differentials have already been discussed and are given by the equations

$$\frac{\partial S_{ij}}{\partial T_{k\ell}} = s^D_{ijk\ell}\,; \quad \frac{\partial S_{ij}}{\partial \delta_n} = \frac{\partial E_n}{\partial T_{ij}} = g_{ijn}\,; \quad \frac{\partial E_m}{\partial \delta_n} = 4\pi\beta^T_{mn} \tag{162}$$

where $s^D_{ijk\ell}$ are the elastic compliances of the crystal at constant displacement, g_{ijn} the piezoelectric constants relating strain to electric displacement $/4\pi$, and β^T_{mn} the dielectric "impermeability" tensor measured at constant stress. We designate the partial derivatives

$$\begin{gathered}\frac{\partial S_{ij}}{\partial T_{k\ell}\partial T_{qr}} = N^D_{ijk\ell qr}\,; \quad \frac{\partial^2 S_{ij}}{\partial T_{k\ell}\partial \delta_m} = \frac{\partial^2 E_m}{\partial T_{k\ell}\partial T_{qr}} = M^D_{ijk\ell n} \\ \frac{\partial^2 S_{ij}}{\partial \delta_n \partial \delta_0} = \frac{\partial^2 E_n}{\partial T_{ij}\partial \delta_0} = Q^D_{ijno}\,; \quad \frac{\partial^2 E_m}{\partial \delta_n \partial \delta_0} = O^D_{mno}\end{gathered} \tag{163}$$

The tensors N, M, Q, and O are respectively tensors of rank 6, 5, 4 and 3 whose interpretation is discussed below. Introducing these definitions equations (161) can be written in the form

$$\begin{aligned}S_{ij} &= T_{k\ell}[s^D_{ijk\ell} + \tfrac{1}{2}N^D_{ijk\ell qr}T_{qr} + M^D_{ijk\ell n}\delta_n] + \delta_n[g_{ijn} + \tfrac{1}{2}Q^D_{ijno}\epsilon_0] \\ E_m &= T_{k\ell}[g_{mk\ell} + \tfrac{1}{2}M^D_{ijk\ell n}T_{qr} + Q^D_{k\ell mn}\delta_n] + \delta_n[4\pi\beta^T_{mn} + \tfrac{1}{2}O^D_{mno}\epsilon_0]\end{aligned} \tag{164}$$

Written in this form the interpretation of the second order terms is obvious. $N^D_{ijk\ell qr}$ represents the nonlinear changes in the elastic compliances $s^D_{ijk\ell}$ caused by the application of stress T_{qr}. Since the product of $N_{ijk\ell qr}T_{qr}$ represents a contracted fourth rank tensor, there is a correction term for each elastic compliance. The tensor $M^D_{ijk\ell n}$ can represent either the nonlinear correction to the elastic compliances due to an applied electric displacement D_n or it can represent the correction to the piezoelectric constant g_{ijn} due to the stresses $T_{k\ell}$. By virtue of the second equation of (162), the second equivalence of (163) results. The fourth rank tensor $\frac{1}{2}Q^D_{ijno}$ represents the electrostrictive effect in a crystal for it determines the strains existing in a crystal which are proportional to the square of the electric displacement. Twice the value of the electrostrictive tensor $\frac{1}{2}Q^D_{ijno}$, which appears in the second equation of (164) can be interpreted as the change in the inverse dielectric constant or "impermeability" constant. Since a change in dielectric constant with applied stress causes a double refraction of light through the crystal, this term is the source of the piezo-optical effect in crystals. The third rank tensor O^D_{mno} represents the change in the "impermeability" constant due to an electric field and hence is the source of the electro-optical effect in crystals.

These equations can also be used to discuss the changes that occur in ferroelectric type crystals such as Rochelle Salt when a spontaneous polariza-

tion occurs in the crystal. When spontaneous polarization occurs, the dipoles of the crystal are lined up in one direction in a given domain. For Rochelle salt this direction is the $\pm X$ axis of the crystal. Now the electric displacement D_x is equal to

$$\delta_x = \frac{D_x}{4\pi} = \frac{E_x}{4\pi} + P_{x_0} + P_{x_D} = \frac{E_x \epsilon_0}{4\pi} P_{x_D} \tag{165}$$

where P_{x_0} is the electronic and atomic polarization, and P_{x_D} the dipole polarization The electronic and atomic polarization is determined by the field and hence can be combined with the field through the dielectric constant ϵ_0, which is the temperature independent part of the dielectric constant. When the crystal becomes spontaneously polarized, a field E_x will result, but this soon is neutralized by the flow of electrons through the surface and volume conductance of the crystal and in a short time $E_x = 0$. Hence for any permanent changes occurring in the crystal we can set

$$\delta_x = \frac{D_x}{4\pi} = P_{x_D} = \text{dipole polarization} \tag{166}$$

which we will write hereafter as P_1.

In the absence of external stresses the direct effects of spontaneous polarization are a spontaneous set of strains introduced by the product of the spontaneous polarization by the piezoelectric constant, and another set produced by the square of the polarization times the appropriate electrostrictive components. For example, Rochelle salt has a spontaneous polarization P_1 along the X_1 axis between the temperatures $-18°$C to $+24°$C. The curve for the spontaneous polarization as a function of temperature is shown by Fig. 6.[6] The only piezoelectric constant causing a spontaneous strain will be $g_{14/2} = g_{123}$. Hence the spontaneous polarization causes a spontaneous shearing strain

$$S_4 = g_{14} P_x = 120 \times 10^{-8} \times 760 = 9.1 \times 10^{-4} \tag{167}$$

if we introduce the experimentally determined values. Since S_4 is the cosine of 90° plus the angle of distortion, this would indicate that the right angled axes of a rhombic system would be distorted 3.1 minutes of arc. This is the value that should hold for any domain. For a crystal with several domains, the resulting distortion will be partly annulled by the different signs of the polarization and should be smaller. Mueller[7] measured an angle of 3′45″ at 0°C for one crystal. This question has also been

[6] This has been measured by measuring the remanent polarization, when all the domains are lined up. See "The Dielectric Anomalies of Rochelle Salt," H. Mueller, *Annals of the N. Y. Acad. Science*, Vol. XL, Art. 5, page 338, Dec. 31, 1940.

[7] "Properties of Rochelle Salt," H. Mueller, *Phys. Rev.*, Vol. 57, No. 9, May 1, 1940.

investigated by the writer and Miss E. J. Armstrong by measuring the temperature expansion coefficients of the Y and Z axes and comparing their average with the expansion coefficient at 45° from these two axes. The difference between these two expansion coefficients measures the change in angle between the Y and Z axes caused by the spontaneous shearing strains. The results are shown by Fig. 7. Above and below the ferroelectric region, the expansion of the 45° crystal coincides with the average expansion of the Y and Z axes measured from 25°C as a reference temperature. Between the Curie temperatures a difference occurs indicating that the Y and Z crystallographic axes are no longer at right angles. The difference in expansion per unit length at 0°C (the maximum point) corresponds to 1.6×10^{-4} cm per cm. This represents an axis distortion of 1.1 minutes

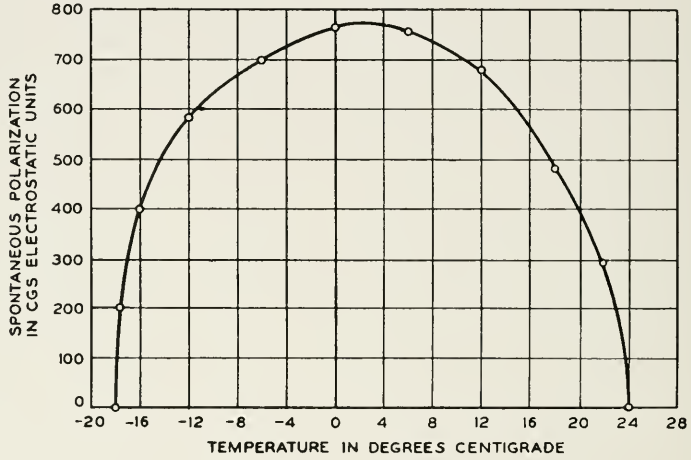


Fig. 6.—Spontaneous polarization in Rochelle Salt along the X axis.

of arc. Correspondingly smaller values are found at other temperatures in agreement with the smaller spontaneous polarization at other temperatures. It was also found that practically the same curve resulted for either 45° axis, indicating that the mechanical bias put on by the optometer used for measuring expansions introduced a bias determining the direction of the largest number of domains.

The second order terms caused by the square of the spontaneous polarization is given by the expression

$$S_{ij} = Q^{D}_{ij11} P_1^2 \tag{168}$$

Since Q is a fourth rank tensor the possible terms for an orthorhombic bisphenoidal crystal (the class for Rochelle salt) are

$$S_{11} = Q^{D}_{1111} P_1^2 ; \qquad S_{22} = Q^{D}_{2211} P_1^2 ; \qquad S_{33} = Q^{D}_{3311} P_1^2 \tag{169}$$

In an effort to measure these effects, careful measurements have been made of the temperature expansions of the three axes X, Y and Z. The results are shown by Table II. On account of the small change in dimension from

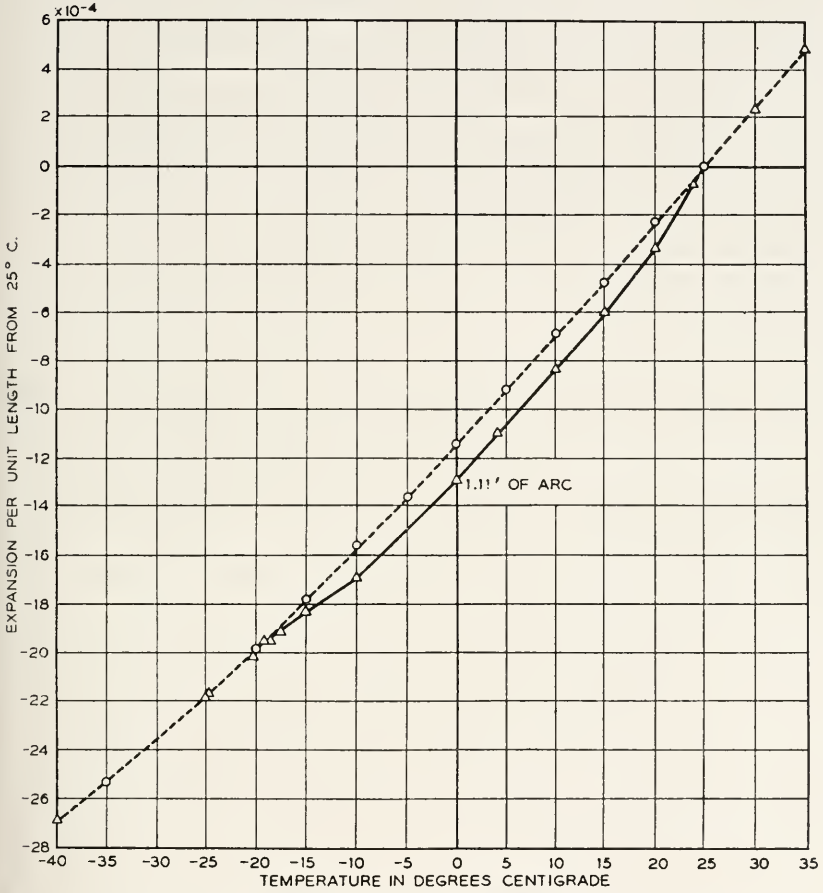


Fig. 7.—Temperature expansion curve along an axis 45° between Y and Z as a function of temperature.

the standard curve it is difficult to pick out the spontaneous components by plotting a curve. By expressing the expansion in the form of the equation

$$\frac{\Delta L}{L} = a_1(T\text{-}25) + a_2(T\text{-}25)^2 + a_3(T\text{-}25)^3 \qquad (170)$$

TABLE II

MEASURED TEMPERATURE EXPANSIONS FOR THE THREE CRYSTALOGRAPHIC AXES

Temperature	Expansion		Expansion		Expansion
in °C.	$\times 10^{-4}$ X Axis	Temperature in °C.	$\times 10^{-4}$ Y Axis	Temperature in °C.	$\times 10^{-4}$ Z Axis
39.6	10.2	+35.0	4.45	+34.9	+4.9
38.7	9.46	30.3	2.5	29.9	2.5
35.2	6.96	25.25	0.2	25.05	+.05
30.2	3.63	23.9	−0.42	24.0	−.5
27.2	1.41	22.9	−0.88	19.95	−2.62
26.2	0.71	19.35	−2.4	14.95	−5.11
25.15	0.06	14.9	−4.25	+9.75	−7.55
24.0	−0.71	10.0	−6.25	+4.9	−9.9
23.0	−1.39	5.4	−8.18	0	−12.31
21.8	−2.37	+0.3	−10.15	−6.35	−15.3
16.0	−6.5	−9.7	−13.98	−10.5	−17.29
15.2	−7.05	−16.3	−16.41	−15.0	−19.42
4.9	−14.12	−20.85	−17.94	−18.0	−20.86
+0.3	−17.28	−25.1	−19.22	−23.2	−23.08
−4.7	−20.3	−30.3	−20.8	−25.1	−23.96
−10.7	−24.0	−35.0	−22.23	−31.1	−26.59
−15.3	−26.6	−39.7	−23.54	−35.0	−28.28
−20.7	−30.2	−53.2	−27.60	−40.0	−30.4
−25.7	−32.7				
−30.1	−35.2				
−34.7	−37.85				
−40.7	−41.25				
−45.0	−44.0				
−50.5	−47.0				

and evaluating the constants by employing temperatures outside of the ferroelectric range, a normal curve was established. For the X, Y, and Z axes these relations are

$$\frac{\Delta L}{L} = 69.6 \times 10^{-6}(T\text{-}25) + 7.4 \times 10^{-8}(\text{T-}25)^2 - 3.13 \times 10^{-10}(\text{T-}25)^3 \quad (X \text{ direction})$$

$$\frac{\Delta L}{L} = 43.7 \times 10^{-6}(\text{T-}25) + 8.16 \times 10^{-8}(\text{T-}25)^2 - 3.60 \times 10^{-10}(\text{T-}25)^3 \quad (Y \text{ direction}) \qquad (171)$$

$$\frac{\Delta L}{L} = 49.4 \times 10^{-6}(T\text{-}25) + 1.555 \times 10^{-8}(T\text{-}25)^2 - 2.34 \times 10^{-10}(T\text{-}25)^3 \quad (Z \text{ direction})$$

The difference between the normal curves and the measured values in the Curie region is shown plotted by the points of Fig. 8. The solid and dashed curves represent curves proportional to the square of the spontaneous polarization and with multiplying constants adjusted to give the best fits for the measured points. These give values of Q^D_{1111}, Q^D_{2211}, Q^D_{3311} equal to

$$Q^D_{1111} = -86.5 \times 10^{-12}; \qquad Q^D_{2211} = +17.3 \times 10^{-12}; \qquad Q^D_{3311} = -24.2 \times 10^{-12} \quad (172) \tag{172}$$

Another effect noted for Rochelle salt is that some of the elastic constants suddenly change by small amounts at the Curie temperatures. This is a consequence of the tensor $M^D_{ijk\ell_n}$, for if a spontaneous polarization P

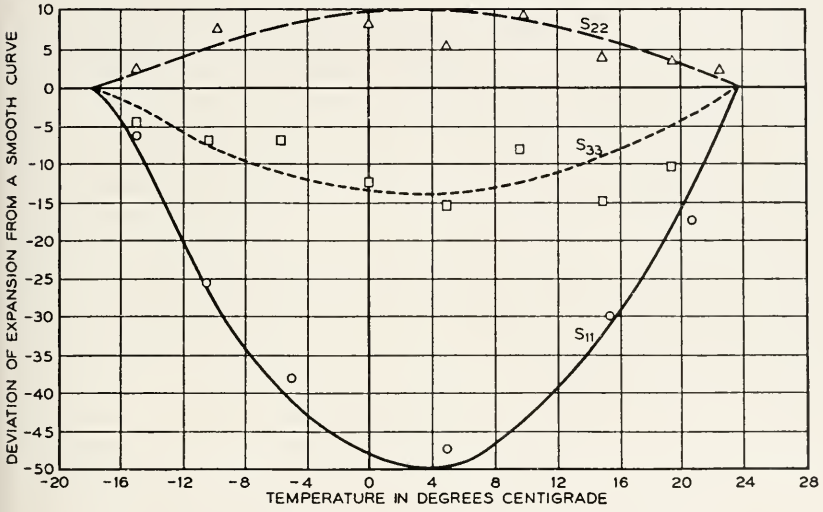


Fig. 8.—Spontaneous electrostrictive strain in Rochelle Salt along the three crystallographic axes.

occurs, a sudden change occurs in some of the elastic constants as can be seen from the first of equations (164). The second equation of (164) shows that this same tensor causes a nonlinear response in the piezoelectric constant. Since a change in the elastic constant is much more easily determined than a nonlinear change in the piezoelectric constant, the first effect is the only one definitely determined experimentally. Since all three crystallographic axes are binary axes in Rochelle salt, it is easily shown that the only terms that can exist for a fifth rank tensor are terms of the types

$$M^D_{11123}\ ; \quad M^D_{12223}\ ; \quad M^D_{12333} \tag{173}$$

with permutations and combinations of the indices. Hence when a spontaneous polarization P_1 occurs, the elastic constants become

$$s^D_{ijk\ell} - M^D_{ijk\ell_1} P_1 \tag{174}$$

Comparing these with the relation of (90) we see that the spontaneous polarization has added the elastic constants

$$
\begin{aligned}
s^D_{14} &= \frac{(M^D_{11231} + M^D_{11321} + M^D_{23111} + M^D_{32111})P_1}{2} \\
s^D_{24} &= \frac{(M^D_{22231} + M^D_{22321} + M^D_{23221} + M^D_{32221})P_1}{2} \\
s^D_{34} &= \frac{(M^D_{23331} + M^D_{32331} + M^D_{33321} + M^D_{33231})P_1}{2} \\
s^D_{56} &= \frac{(M^D_{12151} + M^D_{13211} + M^D_{31121} + M^D_{31211} + M^D_{12131} + M^D_{12311} + M^D_{21131} + M^D_{21311})P_1}{2}
\end{aligned}
\tag{175}
$$

between the two Curie points. Hence while the spontaneous polarization P_1 exists, the resulting elastic constants are

$$
\begin{vmatrix}
s_{11}, & s_{12}, & s_{13}, & s_{14}, & 0, & 0 \\
s_{12}, & s_{22}, & s_{23}, & s_{24}, & 0, & 0 \\
s_{13}, & s_{23}, & s_{33}, & s_{34}, & 0, & 0 \\
s_{14}, & s_{24}, & s_{34}, & s_{44}, & 0, & 0 \\
0, & 0, & 0, & 0, & s_{55}, & s_{56} \\
0, & 0, & 0, & 0, & s_{56}, & s_{66}
\end{vmatrix}
\tag{176}
$$

Comparing this to equation (139) which shows the possible elastic constants for the various crystal classes, we see that between the two Curie points, the crystal is equivalent to a monoclinic sphenoidal crystal (Class 3) with the X axis the binary axis. Outside the Curie region the crystal becomes orthorhombic bisphenoidal. This interpretation agrees with the temperature expansion curves of Fig. 7.

The sudden appearance of the polarization P_1 will affect the frequency of a 45° X-cut crystal, for with a crystal cut normal to the X axis and with the length of the crystal at an angle Θ with the Y axis, the value of the elastic compliance s'_{22} along the length is

$$
\begin{aligned}
s^{D\prime}_{22} = {} & s^D_{22}\cos^4\Theta + 2s^D_{24}\cos^3\Theta\sin\Theta + (2s^D_{23} + s^D_{44})\sin^2\Theta\cos^2\Theta \\
& + 2s^D_{34}\sin^3\Theta\cos\Theta + s^D_{33}\sin^4\Theta
\end{aligned}
\tag{177}
$$

Hence for a crystal with its length 45° between the Y and Z axes, elastic compliance becomes

$$
s'^D_{22} = \frac{s^D_{22} + 2(s^D_{24} + s^D_{23} + s^D_{34}) + s^D_{44} + s^D_{33}}{4} \tag{178}
$$

For a 45° X-cut crystal we would expect a sudden change in the value of s'^{D}_{22} as the crystal becomes spontaneously polarized between the two Curie points due to the addition of the s^{D}_{24} and s^{D}_{34} elastic compliances. Such a change has been observed for Rochelle salt[8] as shown by Fig. 9 which shows the frequency constant of a nonplated crystal for which the elastic compliances s^{D}_{ij} should hold.

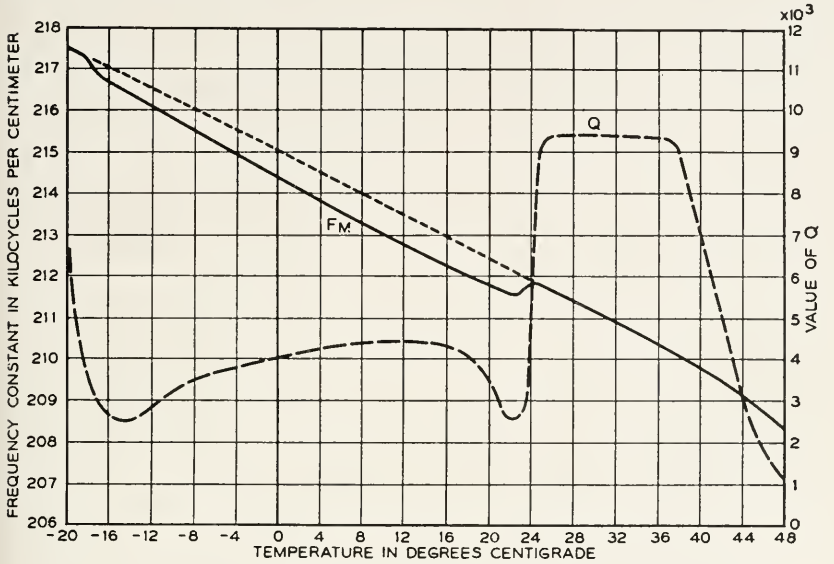


Fig. 9.—Frequency constant and Q of an unplated 45°X cut Rochelle Salt crystal plotted as a function of temperature.

Hence the sudden change in the elastic constant is a result of the two second order terms $s^{D}_{24} + s^{D}_{34}$, which are caused by the spontaneous polarization. The value of the sum of these two terms at the mean temperature of the Curie range, 3°C is

$$s^{D}_{24} + s^{D}_{34} = 4.1 \times 10^{-14}\ \text{cm}^2/\text{dyne} \tag{179}$$

Crystals cut normal to the Y and Z axes should not show a spontaneous change in their frequency characteristic since the spontaneous terms s_{14}, s_{24}, s_{34} and s_{56} do not affect the value of Young's modulii in planes normal to Y and Z. Experiments on a 45° Y-cut Rochelle salt crystal do not show a spontaneous change in frequency at the Curie temperature, although there is a large change in the temperature coefficient of the elastic compliance between the two Curie points. This is the result of third order term and is

[8] "The Location of Hysteresis Phenomena in Rochelle Salt Crystals," W. P. Mason, *Phys. Rev.*, Vol. 50, p. 744–750, October 15, 1940.

not considered here. The spontaneous s_{56} constant affects the shear constant s'_{66} for crystals rotated about the X axis and could be detected experimentally. No experimental values have been obtained.

The effects of spontaneous polarization in the second equation of (164) are of two sorts. For an unplated crystal, a spontaneous voltage is generated on the surface, which, however, quickly leaks off due to the surface and volume leakage of the crystal. The other effects are that the spontaneous polarization introduces new piezoelectric constants through the tensor Q^D_{kfmn}, changes the dielectric constants through the tensor O^D_{mno} and introduces a stress effect on the piezoelectric constants through the tensor M^D_{kfmqr}. Since piezoelectric constants are not as accurately measured as elastic constants, the first effect has not been observed. The additional piezoelectric constants introduced by the tensor Q^D_{kfmn} are shown by equation (180)

$$\begin{matrix} g_{11} & g_{12} & g_{13} & g_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & g_{25} & g_{26} \\ 0 & 0 & 0 & 0 & g_{35} & g_{36} \end{matrix} \tag{180}$$

Since the only constants for the Rochelle salt class, the orthorhombic bisphenoidal, are g_{14}, g_{25}, g_{36}, this shows that between the two Curie points the crystal becomes monoclinic sphenoidal, with the X axis being the binary axis. The added constants are, however, so small that the accuracy of measurement is not sufficient to evaluate them. From the expansion measurements of equation (172) and the spontaneous polarization values, three of them should have maximum values of

$$g_{11} = -6.6 \times 10^{-8}; \qquad g_{12} = +1.3 \times 10^{-8}; \qquad g_{13} = -1.8 \times 10^{-8} \tag{181}$$

These amount to only 6 per cent of the constant g_{14}, and hence they are not easily evaluated from piezoelectric measurements.

The effect of the tensor O^D_{mno} is to introduce a spontaneous dielectric constant ϵ_{23} between the Curie temperatures so that the dielectric tensor becomes

$$\begin{matrix} \epsilon_{11}, & 0, & 0 \\ 0, & \epsilon_{22}, & \epsilon_{23} \\ 0, & \epsilon_{23}, & \epsilon_{33} \end{matrix} \tag{182}$$

As discussed at length by Mueller[9] this introduces a spontaneous birefringence for light passing through the crystal along the X, Y and Z axes which adds to the birefringence already present.

[9] "Properties of Rochelle Salt I and IV," *Phys. Rev.* 47, 175 (1935); 58, 805 November 1, 1940.

The Biased Ideal Rectifier

By W. R. BENNETT

Methods of solution and specific results are given for the spectrum of the response of devices which have sharply defined transitions between conducting and non-conducting regions in their characteristics. The input wave consists of one or more sinusoidal components and the operating point is adjusted by bias, which may either be independently applied or produced by the rectified output itself.

Introduction

THE concept of an ideal rectifier gives a useful approximation for the analysis of many kinds of communication circuits. An ideal rectifier conducts in only one direction, and by use of a suitable bias may have the critical value of input separating non-conduction from conduction shifted to any arbitrary value, as illustrated in Fig. 1. A curve similar to Fig. 1 might represent for example the current versus voltage relation of a biased diode. By superposing appropriate rectifying and linear characteristics with different conducting directions and values of bias, we may approximate the characteristic of an ideal limiter, Fig. 2, which gives constant response when the input voltage falls outside a given range. Such a curve might approximate the relationship between flux and magnetizing force in certain ferromagnetic materials, or the output current versus signal voltage in a negative-feedback amplifier. The abrupt transitions from non-conducting to conducting regions shown are not realizable in physical circuits, but the actual characteristics obtained in many devices are much sharper than can be represented adequately by a small number of terms in a power series or in fact by any very simple analytic function expressible in a reasonably small number of terms valid for both the non-conducting and conducting regions.

In the typical communication problem the input is a signal which may be expressed in terms of one or more sinusoidal components. The output of the rectifier consists of modified segments of the original resultant of the individual components separated by regions in which the wave is zero or constant. We are not so much interested in the actual wave form of these chopped-up portions, which would be very easy to compute, as in the frequency spectrum. The reason for this is that the rectifier or limiter is usually followed by a frequency-selective circuit, which delivers a smoothly varying function of time. Knowing the spectrum of the chopped input to the selective network and the steady-state response as a function of

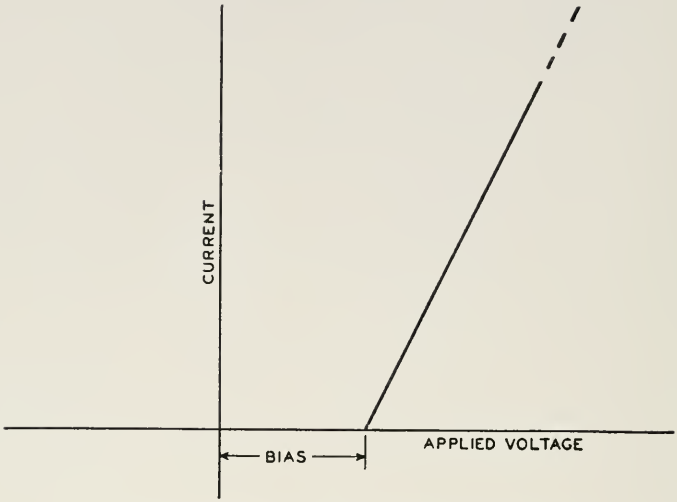


Fig. 1.—Ideal biased linear rectifier characteristic.

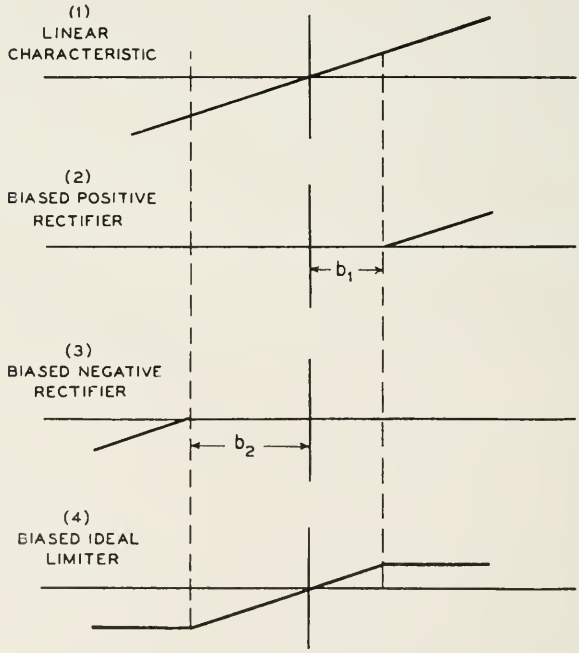


Fig. 2.—Synthesis of limiter characteristic.

frequency of the network, we can calculate the output wave, which is the one having most practical importance. The frequency selectivity may in many cases be an inherent part of the rectifying or limiting action so that discrete separation of the non-linear and linear features may not actually be possible, but even then independent treatment of the two processes often yields valuable information.

The formulation of the analytical problem is very simple. The standard theory of Fourier series may be used to obtain expressions for the amplitudes of the harmonics in the rectifier output in the case of a single applied frequency, or for the amplitudes of combination tones in the output when two or more frequencies are applied. These expressions are definite integrals involving nothing more complicated than trigonometric functions and the functions defining the conducting law of the rectifier. If we were content to make calculations from these integrals directly by numerical or mechanical methods, the complete solutions could readily be written down for a variety of cases covering most communication needs, and straightforward though often laborious computations could then be based on these to accumulate eventually a sufficient volume of data to make further calculations unnecessary.

Such a procedure however falls short of being satisfactory to those who would like to know more about the functions defined by these integrals without making extensive numerical calculations. A question of considerable interest is that of determining under what conditions the integrals may be evaluated in terms of tabulated functions or in terms of any other functions about which something is already known. Information of this sort would at least save numerical computing and could be a valuable aid in studying the more general aspects of the communication system of which the rectifier may be only one part. It is the purpose of this paper to present some of these relationships that have been worked out over a considerable period of time. These results have been found useful in a variety of problems, such as distortion and cross-modulation in overloaded amplifiers, the performance of modulators and detectors, and effects of saturation in magnetic materials. It is hoped that their publication will not only make them available to more people, but also stimulate further investigations of the functions encountered in biased rectifier problems.

The general forms of the integrals defining the amplitudes of harmonics and side frequencies when one or two frequencies are applied to a biased rectifier are written down in Section I. These results are based on the standard theory of Fourier series in one or more variables. Some general relationships between positive and negative bias, and between limiters and biased rectifiers are also set down for further reference. Some discussion is given of the modifications necessary when reactive elements are used in the circuit.

Section II summarizes specific results on the single-frequency biased rectifier case. The general expression for the amplitude of the typical harmonic is evaluated in terms of a hypergeometric function for the power law case with arbitrary exponent.

Section III takes up the evaluation of the two-frequency modulation products. It is found that the integer-power-law case can be expressed in finite form in terms of complete elliptic integrals of the first, second, and third kind for almost all products. Of these the first two are available in tables, directly, and the third can be expressed in terms of incomplete integrals of the first and second kinds, of which tables also exist. No direct tabulation of the complete elliptic integrals of the third kind encountered here is known to the author. They are of the hyperbolic type in contrast to the circular ones more usual in dynamical problems. Imaginary values of the angle β would be required in the recently published table by Heuman[1].

A few of the product amplitudes depend on an integral which has not been reduced to elliptic form, and which is a transcendental function of two variables about which little is known. Graphs calculated by numerical integration are included.

The expressions in terms of elliptic integrals, while finite for any product, show a rather disturbing complexity when compared with the original integrals from which they are derived. It appears that elliptic functions are not the most natural ones in which the solution to our problem can be expressed. If we did not have the elliptic tables available, we would prefer to define new functions from our integrals directly, and the study of such functions might be an interesting and fruitful mathematical exercise.

Solutions for more than two frequencies are theoretically possible by the same methods, although an increase of complexity occurs as the first few components are added. When the number of components becomes very large, however, limiting conditions may be evaluated which reduce the problem to a manageable simplicity again. The case of an infinite number of components uniformly spaced along an appropriate frequency range has been used successfully as a representation of a noise wave, and the detected output from signal and noise inputs thus evaluated[2]. The noise problem will not be treated in the present paper.

I. The General Problem

Let the biased rectifier characteristic, Fig. 1, be expressed by

$$I = \begin{pmatrix} 0, & E < b \\ f(E-b), & b < E \end{pmatrix} \tag{1.1}$$

[1] Carl Heuman, Tables of Complete Elliptic Integrals, *Jour. Math. and Physics*, Vol. XX, No. 2, pp. 127–206, April, 1941.

[2] W. R. Bennett, Response of a Linear Rectifier to Signal and Noise, *Jour. Acous. Soc. Amer.*, Vol. 15, pp. 164–172, Jan. 1944.

Then if a single frequency wave defined by

$$E = P \cos pt, \quad -P < b < P, \tag{1.2}$$

is applied as input, the output contains only the tips of the wave, as shown in Fig. 3. It is convenient to place the restrictions on P and b given in Eq. (1.2). The sign of P is taken as positive since a change of phase may be introduced merely by shifting the origin of time and is of trivial interest. If the bias b were less than $-P$, the complete wave would fall in the conducting region and there would be no rectification. If b were greater than

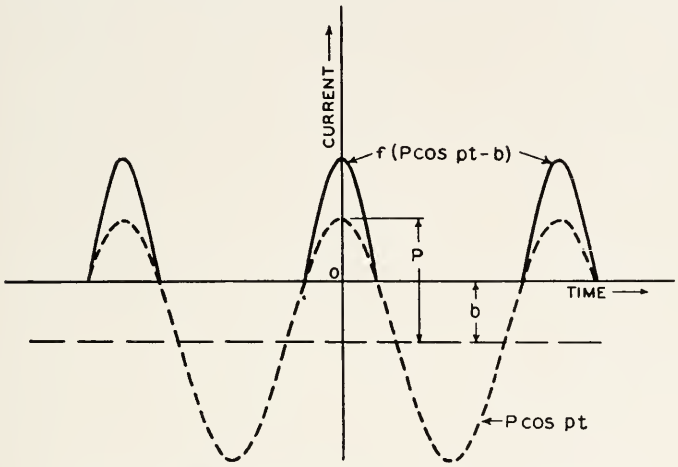


Fig. 3.—Response of biased rectifier to single-frequency wave.

P, the output would be completely suppressed. Applying the theory of Fourier series to (1.1) and (1.2), we have the results

$$I = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos n\, pt \tag{1.3}$$

$$a_n = \frac{2}{\pi} \int_0^{\text{arc cos } b/P} f(P \cos x - b) \cos nx \, dx \tag{1.4}$$

When two frequencies are applied, the output may be represented by a double Fourier series. The typical coefficient may be found by the method explained in an earlier paper by the author[3]. The problem is to obtain the double Fourier series expansion in x and y of the function $g(x, y)$ defined by:

$$g(x, y) = \begin{cases} 0, & P \cos x + Q \cos y < b \\ f(P \cos x + Q \cos y - b), & b < P \cos x + Q \cos y \end{cases} \tag{1.5}$$

[3] W. R. Bennett, New Results in the Calculation of Modulation Products, *B.S.T.J.*, Vol. XII, pp. 228–243, April, 1933.

We substitute the special values $x = pt, y = qt$ after obtaining the expansion. Let

$$k_1 = Q/P, k_0 = -b/P \tag{1.6}$$

The most general conditions of interest are comprised in the ranges:

$$0 < k_1 < 1, \quad -2 < k_0 < 2 \tag{1.7}$$

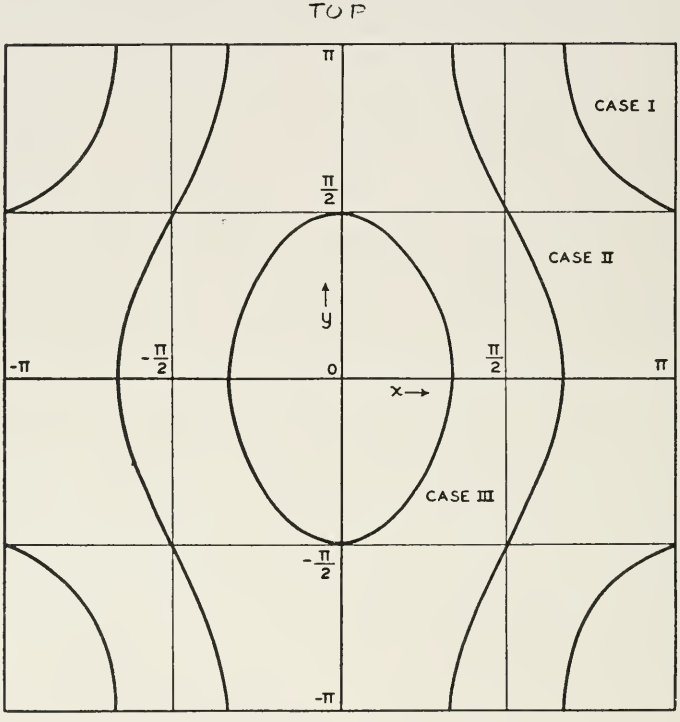


Fig. 4.—Regions in xy-plane bounded by $k_0 + \cos x + k_1 \cos y = 0$.

The regions in the xy-plane in which $g(x,y)$ does not vanish are bounded by the various branches of the curve:

$$k_0 + \cos x + k_1 \cos y = 0 \tag{1.8}$$

We need to consider only one period rectangle bounded by $x = \pm\pi, y = \pm\pi$, since the function repeats itself at intervals of 2π in both x and y. The shape of the curve (1.8) within this rectangle may have three forms, which are depicted in Fig. 4. In Case I, $k_0 + k_1 > k$, $k_0 - k_1 < 1$, the curve divides into four branches which are open at both ends of the x- and y-axes. In Case (2), $k_0 + k_1 < 1$, $k_0 - k_1 > -1$, the curve has two branches open

at the ends of the y-axis. In Case (3), $-1 < k_0 + k_1 < 1$, $k_0 - k_1 < -1$, a single closed curve is obtained. The limits of integration must be chosen to fit the proper case. The Fourier series expansion of $g(x,y)$ may be written:

$$g(x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} \cos mx \cos ny \tag{1.9}$$

where a_{mn} is found from integrals of the form:

$$A = \frac{\epsilon_m \epsilon_n}{\pi^2} \int_{y_1}^{y_2} dy \int_{x_1}^{x_2} f(P \cos x + Q \cos y - b) \cos mx \cos ny \, dx \tag{1.10}$$

Here, as usual, ϵ_m is Neumann's discontinuous factor equal to two when m is not zero and unity when m is zero. The values of the limits for the different cases are:

Case I, $a_{mn} = A_1 + A_2$

$$A_1 = A \text{ with limits} \begin{pmatrix} (x_1 = 0, & x_2 = \arccos(-k_0 - k_1 \cos y) \\ y_1 = \arccos \dfrac{1 - k_0}{k_1}, & y_2 = \pi \end{pmatrix} \tag{1.11}$$

$$A_2 = A \text{ with limits} \begin{pmatrix} x_1 = 0, & x_2 = \pi \\ y_1 = 0, & y_2 = \arccos \dfrac{1 - k_0}{k_1} \end{pmatrix} \tag{1.12}$$

Case II, $a_{mn} = A$

$$\text{Limits} \begin{pmatrix} x_1 = 0, & x_2 = \arccos(-k_0 - k_1 \cos y \\ y_1 = 0, & y_2 = \pi \end{pmatrix} \tag{1.13}$$

Case III, $a_{mn} = A$

$$\text{Limits} \begin{pmatrix} (x_1 = 0, & x_2 = \arccos(-k_0 - k_1 \cos y) \\ (y_1 = 0, & y_2 = \arccos\left(-\dfrac{1 + k_0}{k_1}\right) \end{pmatrix} \tag{1.14}$$

For a considerable variety of rectifier functions f, the inner integration may be performed at once leaving the final calculation in terms of a single definite integral.

A somewhat different point of view is furnished by evaluating the integral (1.4) for the biased single-frequency harmonic amplitude, and then replacing the bias by a constant plus a sine wave having the second frequency. When each harmonic of the first frequency is in turn expanded in a Fourier series

in the second frequency, the two-frequency modulation coefficients are obtained. Some early calculations carried out graphically in this way are the source of the curves plotted in Figs. 18 to 21 inclusive, for which I am indebted to Dr. E. Peterson.

If reactive elements are used in the rectifier circuit, the voltage across the rectifying element may depart from the input wave shape applied to the complete network. The solution then loses its explicit nature since the rectifier current is expressed in terms of input voltage components which in turn depend on voltage drops produced in the remainder of the network by the rectifier currents. Practical solutions can be worked out when relatively few components are important.

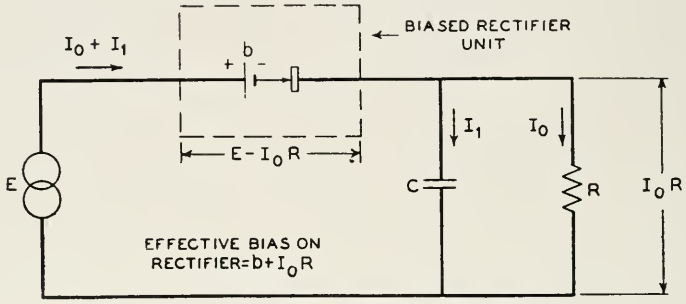


Fig. 5.— Biased rectifier in series with RC network.

As an example consider the familiar case of a parallel combination of resistance R and capacitance C in series with the biased rectifier, Fig. 5. If C has negligible impedance at all frequencies of importance in the rectifier circuit except zero, we may assume that the voltage across R is constant and equal to I_0R, where I_0 is the d-c. component of the rectifier current. The voltage across the rectifier unit is then $E - I_0R$. The effect is a change in the value of bias from b to $b + I_0R$. If the d-c component in the output is calculated for bias $b + I_0R$, we obtain the value of I_0 in terms of $b + I_0R$, an implicit equation defining I_0. If this equation can be solved for I_0, the bias $b + I_0R$ can then be determined and the remaining modulation products calculated.

A more important case is that of the so-called envelope detector, in which the impedance of the condenser is very small at all frequencies contained in the input signal, but is very large at frequencies comparable with the band width of the spectrum of the input signal. These are the usual conditions prevailing in the detection of audio or video signals from modulated r-f or i-f waves. The solution depends on writing the input signal in the form of a slowly varying positive valued envelope function multiplying a rapidly

oscillating cosine function. That is, if the input signal can be represented as

$$E = A\ (t) \cos \phi\ (t), \tag{1.15}$$

where $A\ (t)$ is never negative and has a spectrum confined to the frequency range in which $2\pi fC$ is negligibly small compared with $1/R$, while $\cos \phi(t)$ has a spectrum confined to the frequency range in which $1/R$ is negligibly small compared with $2\pi fC$, we divide the components in the detector output into two groups, viz.:

1. A low-frequency group I_{lf} containing all the frequencies comparable with those in the spectrum of $A\ (t)$. The components of this group flow through R.

2. A high-frequency group I_{hf} containing all the frequencies comparable to and greater than those in the spectrum of $\cos \phi\ (t)$. The components of this group flow through C and produce no voltage across R.

The instantaneous voltage drop across R is therefore equal to $I_{lf}R$, and hence the bias on the rectifier is $b + I_{lf}R$. If A and ϕ were constants, we could make use of (1.3) and (1.4) to write:

$$I_{lf} + I_{hf} = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos n\theta \tag{1.16}$$

$$a_n = \frac{2}{\pi} \int_0^{\arccos [(b+I_{lf}R)/A]} f(A \cos x - b - I_{lf}R) \cos nx\, dx \tag{1.17}$$

If A and ϕ are variable, the equation still holds provided $I_{lf}R < A$ at all times. Assuming the latter to be true (keeping in mind the necessity of checking the assumption when I_{lf} is found), we note that terms of the form $a_n \cos n\theta$ consist of high frequencies modulated by low frequencies and hence the main portion of their spectra must be in the high-frequency range. Hence we must have as a good approximation when the envelope frequencies are well separated from the intermediate frequencies,

$$I_{lf} = \frac{a_0}{2} = \frac{1}{\pi} \int_0^{\arccos [(b+I_{lf}R)/A]} f(A \cos x - b - I_{lf}R)dx \tag{1.18}$$

This equation defines I_{lf} as a function of A, and if it is found that the condition $b + I_{lf}R < A$ is satisfied by the resulting value of I_{lf}, the problem is solved. If the condition is not satisfied, a more complicated situation exists requiring separate consideration of the regions in which $b + I_{lf}R < A$ and $b + I_{lf}R > A$.

To be specific, consider the case of a linear rectifier with forward conductance $\alpha = 1/R$, and write $V = I_{lf}R$. Then

$$\frac{\pi R_0}{R} V = \sqrt{A - (b + V)^2} - (b + V) \arccos \frac{b + V}{A} \tag{1.19}$$

When $b = 0$ (the case of no added bias), this equation may be satisfied by setting

$$V = cA, 0 \leq c \leq 1, \tag{1.20}$$

which leads to

$$\frac{\pi R_0}{R} = \sqrt{\frac{1}{c^2} - 1} - \arccos c, \tag{1.21}$$

defining c as a function of R_0/R. The value of c approaches unity when the ratio of rectifier resistance to load resistance approaches zero and falls off to zero as R_0/R becomes large. The curve may be found plotted elsewhere[4]. This result justifies the designation of this circuit as an envelope detector since with the proper choice of circuit parameters the output voltage is proportional to the envelope of the input signal.

The equations have been given here in terms of the actual voltage applied to the circuit. The results may also be used when the signal generator contains an internal impedance. For example, a nonreactive source independent of frequency may be combined with the rectifying element to give a new resultant characteristic. If the source impedance is a constant pure resistance r_0 throughout the frequency range of the signal input but is negligibly small at the frequencies of other components of appreciable size flowing in the detector, we assume the voltage drop in r_0 is $r_0 a_1 \cos \phi\ (t)$. We then set $n = 1$ in (1.17) and replace a_1 by $(A_0 - A)/r_0$, where A_0 is the voltage of the source. The value of I_{1f} in terms of A from (1.18) is then substituted, giving an implicit relation between A and A_0.

A further noteworthy fact that may be deduced is the relationship between the envelope and the linearly rectified output. By straightforward Fourier series expansion, the positive lobes of the wave (1.15), may be written as:

$$E_r = \begin{pmatrix} E, & E > 0 \\ 0, & E < 0 \end{pmatrix} = A(t) \left[\frac{1}{\pi} + \frac{1}{2} \cos \phi\ (t) - \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{(-)^m \cos 2m\ \phi(t)}{4m^2 - 1} \right] \tag{1.22}$$

Hence if we represent the low-frequency components of E_r by E_{1f}, we have:

$$E_{1f} = \frac{A(t)}{\pi} \tag{1.23}$$

or

$$A\ (t) = \pi E_{1f} \tag{1.24}$$

[4] See, for example, the top curve of Fig. 9–25, p. 311, H. J. Reich, Theory and Applications of Electron Tubes, McGraw-Hill, 1944.

Equation (1.23) expresses the fact that we may calculate the signal component in the output of a half-wave linear rectifier by taking $1/\pi$ times the envelope. Equation (1.24) shows that we may calculate the response of an envelope detector by taking π times the low-frequency part of the Fourier series expansion of the linearly rectified input. Thus two procedures are in general available for either the envelope detector or linear rectifier solution, and in specific cases a saving of labor is possible by a proper choice between the two methods. The final result is of course the same, although there may be some difficulty in recognizing the equivalence. For example, the solution for linear rectification of a two-frequency wave $P \cos pt + Q \cos qt$ was given by the author in 1933[3], while the solution for the envelope was given by Butterworth in 1929[5]. Comparing the two expressions for the direct-current component, we have:

$\bar{E}_{lf} = \frac{2P}{\pi^2}[2E - (1 - k^2)\,K]$, where K and E are complete elliptic integrals of the first and second kinds with modulus $k = Q/P$

$\overline{A\,(t)} = \frac{2P}{\pi}\,(1 + k)\,E_1$, where E_1 is a complete elliptic intregal of the second kind with modulus $k_1 = 2\sqrt{k}/(1 + k)$. Equation (1.24) implies the existence of the identity

$$(1 + k)\,F_1 = 2E - (1 - k^2)\,K \tag{1.25}$$

The identity can be demonstrated by making use of Landen's transformation in the theory of elliptic integrals.

2. Single-Frequency Signal

The expression for the harmonic amplitudes in the output of the rectifier can be expressed in a particularly compact form when the conducting part of the characteristic can be described by a power law with arbitrary exponent. Thus in (1.4) if $f(z) = \alpha z^{\nu}$, we set $\lambda = b/P$ and get

$$a_n = \frac{2\alpha P^{\nu}}{\pi}\int_0^{\arccos\lambda} (\cos x - \lambda)^{\nu} \cos nx\, dx$$

$$= \frac{2^{\frac{1}{2}}\,\Gamma(\nu + 1)\,aP^{\nu}(1 - \lambda)^{\nu+\frac{1}{2}}}{\pi^{\frac{1}{2}}\,\Gamma\left(\nu + \frac{3}{2}\right)}\,F\left(\frac{1}{2} + n, \frac{1}{2} - n; \nu + \frac{3}{2}; \frac{1 - \lambda}{2}\right) \tag{2.1}$$

[5] S. Butterworth, Apparent Demodulation of a Weak Station by a Stronger One *Experimental Wireless*, Vol. 6, pp. 619–621, Nov. 1929.

The equation holds for all real values of ν greater than -1. The symbol F represents the Gaussian hypergeometric function[6]:

$$F(a, b; c; z) = 1 + \frac{a\,b}{c\,1!}z + \frac{a(a+1)\,b\,(b+1)}{c(c+1)\,2!}z^2 + \cdots \tag{2.2}$$

The derivation of (2.1) requires a rather long succession of substitutions, expansions, and rearrangements, which will be omitted here.

When ν is an integer, the hypergeometric function may be expressed in finite algebraic form, either by performing the integration directly, or by making use of the formulas:

$$\begin{aligned} F(\mu/2, -\mu/2; 1/2; z) &= \cos(\mu \arcsin z), \\ F\left(\frac{1+\mu}{2}, \frac{1-\mu}{2}; \frac{3}{2}; z^2\right) &= \frac{\sin(\mu \arcsin z)}{\mu z}, \end{aligned} \tag{2.3}$$

together with recurrence formulas for the F-function. When ν is an odd multiple of one half, the F-function may be expressed in terms of complete elliptic integrals of the first and second kind with modulus $[(1-\lambda)/2]^{1/2}$ by means of the relations,

$$\left.\begin{aligned} F(\tfrac{1}{2}, \tfrac{1}{2}; 1; k^2) &= \frac{2}{\pi}K, \\ F(-\tfrac{1}{2}, \tfrac{1}{2}; 1; k^2) &= \frac{2}{\pi}E, \end{aligned}\right\} \tag{2.4}$$

and the recurrence formulas for the F-function. For the case of zero bias, we set $\lambda = 0$, and apply the formula

$$F(a, 1-a; c; 1/2) = \frac{\Gamma\left(\frac{c}{2}\right)\Gamma\left(\frac{c+1}{2}\right)}{\Gamma\left(\frac{c+a}{2}\right)\Gamma\left(\frac{1+c-a}{2}\right)} \tag{2.5}$$

obtaining the result:

$$a_n = \frac{2^{\frac{1}{2}}\Gamma(\nu+1)\Gamma\left(\frac{2\nu+3}{4}\right)\Gamma\left(\frac{2\nu+5}{4}\right)\alpha P^\nu}{\pi^{\frac{1}{2}}\Gamma(\nu+3/2)\Gamma\left(\frac{2+\nu+n}{2}\right)\Gamma\left(\frac{2+\nu-n}{2}\right)} \tag{2.6}$$

We point out that the above results may be applied not only when the applied signal is of the form $P\cos pt$ with P and p constants, but to signals

[6] For an account of the properties of the hypergeometric function, see Ch. XIV of Whittaker and Watson, Modern Analysis, Cambridge, 1940. A discussion of elliptic integrals is given in Ch. XXII of the same book.

in which P and p are variable, provided that P is always positive. We thus can apply the results to detection of an ordinary amplitude-modulated wave or to the detection of a frequency-modulated wave after it has passed through a slope circuit.

A case of considerable practical interest is that of an amplitude-modulated wave detected by a diode in series with a parallel combination of resistance R and capacitance C. The value of C is assumed to be sufficiently large so that the voltage across R is equal to the $a_0/2$ component of the current through the diode multiplied by the resistance. This is the condition for envelope detection mentioned in Part 1. The diode is assumed to follow Child's law, which gives $\nu = 3/2$. We write

$$I_0 = \frac{V}{R} = a_0/2 = \frac{\Gamma(5/2)(1-\lambda)^2 \alpha P^{3/2}}{(2\pi)^{\frac{1}{2}}\Gamma(3)} F\left(\tfrac{1}{2}, \tfrac{1}{2}; 3; \frac{1-\lambda}{2}\right) \quad (2.7)$$

where $\lambda = V/P$. Note that V is a constant equal to the direct-voltage output if P is constant. If P varies slowly with time compared with the high-frequency term $\cos pt$, V represents the slowly varying component of the output and hence is the recovered signal.
But

$$F(\tfrac{1}{2}, \tfrac{1}{2}; 3; k^2) = \frac{16}{9\pi k^4}[2(2k^2 - 1)E + (2 - 3k^2)(1 - k^2)K] \quad (2.8)$$

where K and E are complete elliptic integrals of the first and second kind with modulus k. Hence

$$\frac{3\pi}{R_1\sqrt{2P}} = \rho = \frac{(1 + 3\lambda)(1 + \lambda)}{\lambda} K - 8E \quad (2.9)$$

where the modulus of K and E is $\sqrt{(1-\lambda)/2}$. This equation defines ρ as a function of λ, and hence by inversion gives λ as a function of ρ. The resulting curve of λ vs. ρ is plotted in Fig. 6 and may be designated as the function $\lambda = g(\rho)$. If we substitute $\lambda = V/P$ we then have

$$V = P\, g(3\pi/R\alpha\sqrt{2P}) \quad (2.10)$$

This enables us to plot V as a function of P, for various values of $R\alpha$, Fig. 7. Since P may represent the envelope of an amplitude-modulated (or differentiated FM) wave, and V the corresponding recovered signal output voltage, the curves of Fig. 7 give the complete performance of the circuit as an envelope detector. In general the envelope would be of form $P = P_0[1 + c\, s(t)]$, where $s(t)$ is the signal. We may substitute this value of P directly in (2.10) provided the absolute value of $c\, s(t)$ never exceeds unity.

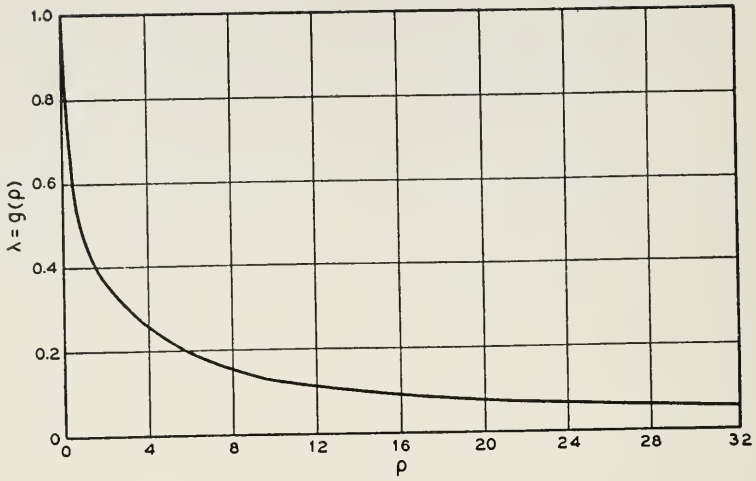


Fig. 6.—The Function $\lambda = g(\rho)$ defined by Eq. (2.9).

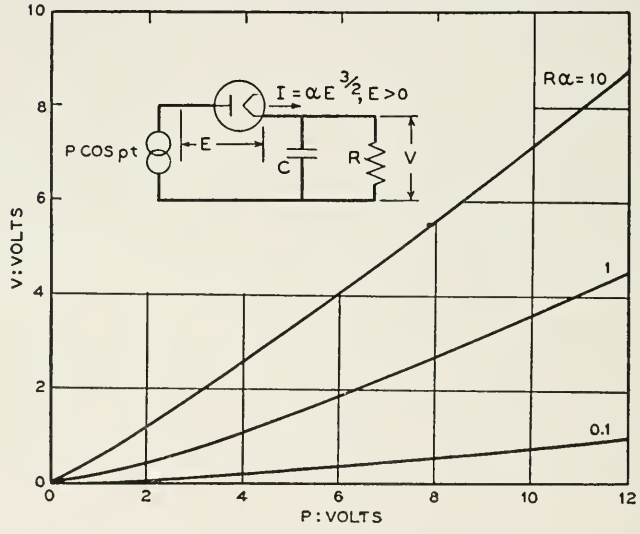


Fig. 7.—Performance of 3/2—power-law rectifier as an envelope detector with low-impedance signal generator.

To express the output in terms of a source voltage P_0 in series with an impedance equal to the real constant value r_0 at the signal frequency and zero at all other frequencies, we write

$$\frac{P_0 - P}{r_0} = a_1 = \frac{3\alpha P^{3/2}(1 - \lambda)^2}{4\sqrt{2}} F\left(\tfrac{3}{2}, -\tfrac{1}{2}; 3; \frac{1 - \lambda}{2}\right) \qquad (2.11)$$

or

$$P_0 = \left(1 + \frac{r_0}{R} H\right) P, \tag{2.12}$$

where

$$H = \frac{3R\alpha(1-\lambda)^2 p^{\frac{1}{2}}}{4\sqrt{2}} F\left(\tfrac{3}{2}, -\tfrac{1}{2}; 3; \frac{1-\lambda}{2}\right)$$
$$= \frac{8R\alpha}{5\pi}\sqrt{2P}[2(1 - k^2 + k^4)E - (2 - k^2)(1 - k^2)K]. \tag{2.13}$$

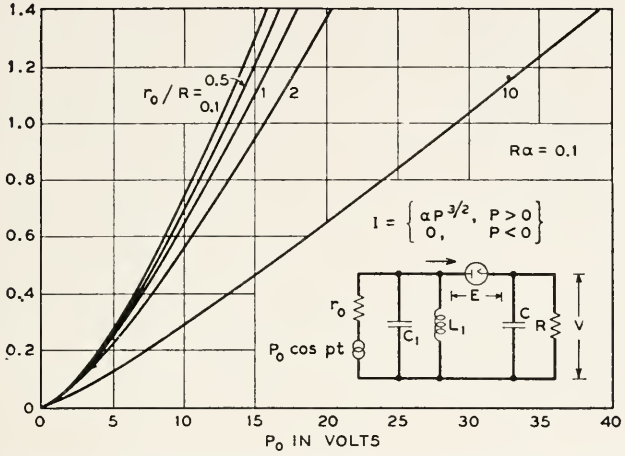


Fig. 8.—Performance of 3/2—power-law rectifier as an envelope detector with impedance of signal generator low except in signal band.

By combining the curves of Fig. 7 giving V in terms of P with the above equations giving the relation between P and P_0, we obtain the curves of Figs. 8, 9, 10, giving V as a function of P_0. The curves approach linearity as $R\alpha$ is made large. On the assumption that the curves are actually linear, we define the conversion loss D of the detector in *db* in terms of the ratio of maximum power available from the source to the power delivered to the load:

$$D = 10 \log_{10} \frac{P_0^2/8r_0}{V^2/R} = 10 \log_{10} \left(\frac{P_0}{V}\right)^2 \frac{R}{8r_0}. \tag{2.14}$$

Curves of D vs r_0/R are given in Figs. 11 and 12. The optimum relation between r_0 and R when the forward resistance of the rectifier vanishes has long been known to be $r_0/R = .5$. The curves show a minimum in this

region when $R\alpha$ is large. In the limit as $R\alpha$ approaches infinity, we may show that the relation between P_0 and V approaches:

$$P_0 = V\left(1 + \frac{2r_0}{R}\right) \tag{2.15}$$

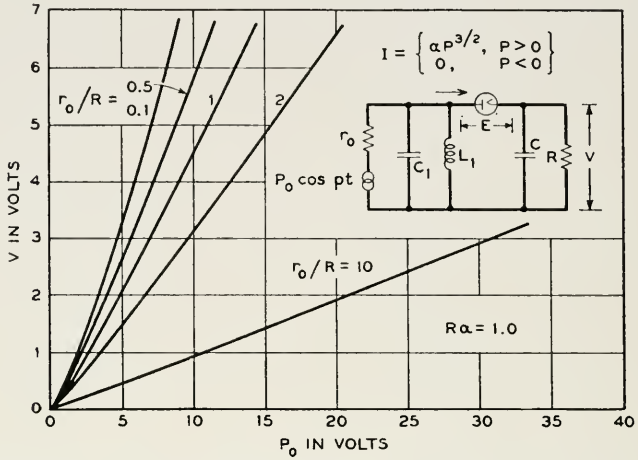


Fig. 9.—Performance of 3/2—power-law rectifier as an envelope detector with impedance of signal generator low except in signal band.

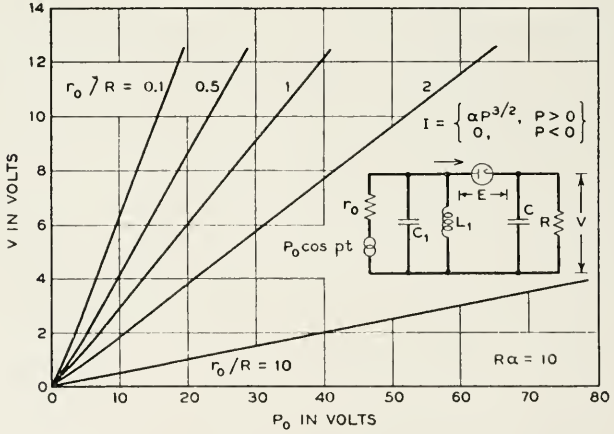


Fig. 10.—Performance of 3/2—power-law rectifier as an envelope detector with impedance of signal generator low except in signal band.

The corresponding limiting formula for D is

$$D = 10 \log_{10} \frac{R}{8r_0}\left(1 + \frac{2r_0}{R}\right)^2 \tag{2.16}$$

The minimum value of D is then found to occur at $r_0 = R/2$ and is zero *db*. We note from the curves that the minimum loss is 1.2 *db* when $R\alpha = 10$ and 0.4 when $R\alpha = 100$.

This example is intended mainly as illustrative rather than as a complete tabulation of possible detector solutions. The methods employed are sufficiently general to solve a wide variety of problems, and the specific evaluation process included should be sufficiently indicative of the procedures required. Cases in which various other selective networks are associated with the detector have been treated by Wheeler[7].

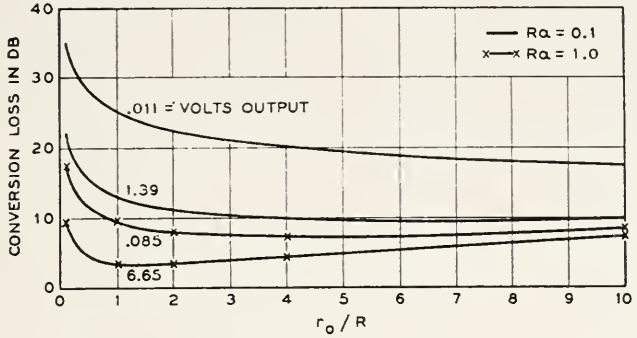


Fig. 11.—Conversion loss of 3/2—power-law rectifier as envelope detector with impedance of signal generator low except in signal band.

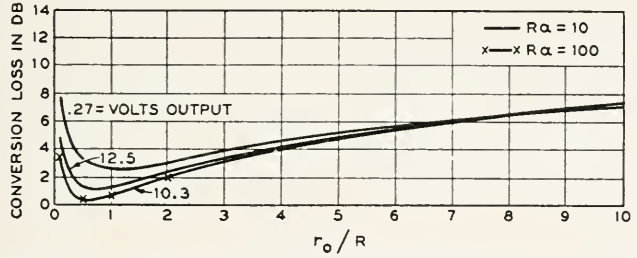


Fig. 12.—Conversion loss of 3/2—power-law rectifier as envelope detector with impedance of signal generator low except in signal band.

3. Two-Frequency Inputs

The general formula for the coefficients in the two-frequency case depends on a double integral as indicated by (1.10). In many cases one integration may be performed immediately, thereby reducing the problem to a single definite integral which may readily be evaluated by numerical or mechanical

[7] H. A. Wheeler, Design Formulas for Diode Detectors, *Proc. I. R. E.*, Vol. 26, pp. 745–780, June 1938.

means. It appears likely in most cases that the expression of these results in terms of a single integral is the most advantageous form for practical purposes, since the integrands are relatively simple, while evaluations in terms of tabulated functions, where possible, often lead to complicated terms. Numerical evaluation of the double integral is also a possible method in cases where neither integration can be performed in terms of functions suitable for calculation.

One integration can always be accomplished for the integer power-law case, since the function $f\ (P \cos x + Q \cos y - b)$ in (1.12) then becomes a polynomial in $\cos x$ and $\cos y$. Cases of most practical interest are the zero-power, linear, and square-law detectors, in which $f(z)$ is proportional to z^0, z^1, and z^2 respectively. The zero-power-law rectifier is also called a total limiter, since it limits on infinitesimally small amplitudes. We shall tabulate here the definite integrals for a few of the more important low-order

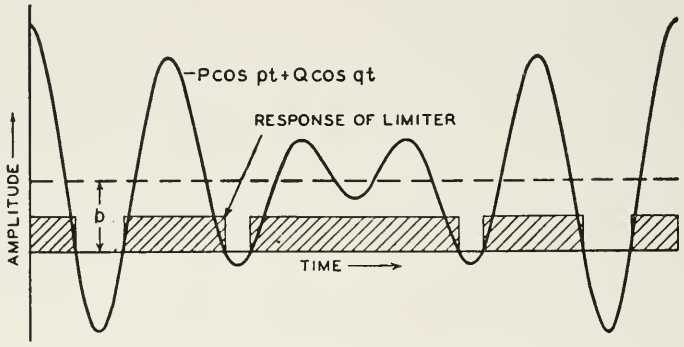


Fig. 13.—Response of biased total limiter to two-frequency wave.

coefficients. To make the listing uniform with that of our earlier work, we express results in terms of the coefficient A_{mn}, which is the amplitude of the component of frequency $mp \pm nq$. The coefficient A_{mn} is half of a_{mn} when neither m nor n is zero. When m or n is zero, we take $A_{mn} = a_{mn}$ and drop the component with the lower value of the $\pm$ sign. When both m and n are zero, we use the designation $A_{00}/2$ for a_{00}, the d-c term. In the tabulations which follow we have set $f(z) = \alpha z^\nu$ with ν taking the values of zero and unity.

We first consider the biased zero-power-law rectifier or biased total limiter. This is the case in which the current switches from zero to a constant value under control of two frequencies and a bias as illustrated by Fig. 13. The results are applicable to saturating devices when the driving forces swing through a large range compared with the width of the linear region. It is also to be noted that the response of a zero-power-law rectifier may be regarded as the Fourier series expansion of the conductance

of a linear rectifier under control of two carrier frequencies and a bias. The results may therefore be applied to general modulator problems based on the method described by Peterson and Hussey[8]. We may also combine the Fourier series with proper multiplying functions to analyze switching between any arbitrary forms of characteristics. We give the results for positive values of k_0. The corresponding coefficients for $-k_0$ can be obtained from the relations:

$$\left.\begin{aligned} \frac{A^{-}_{00}}{2} &= \alpha - \frac{A^{+}_{00}}{2} \\ A^{-}_{mn} &= (-)^{m+n+1} A^{+}_{mn}, \qquad m+n>0 \end{aligned}\right\} \tag{3.1}$$

Here we have used plus and minus signs as superscripts to designate coefficients with bias $+k_0$ and $-k_0$ respectively. We thus obtain a reduction in the number of different cases to consider, since Case III consists of negative bias values only, and these can now be expressed in terms of positive bias values falling in Cases I and II. It is convenient to define an angle θ by the relations:

$$\theta = \begin{pmatrix} \arccos \dfrac{1-k_0}{k_1}, & k_0+k_1>1,\ k_0-k_1<1 \cdot \text{ (Case I)} \\ 0, & k_0+k_1<1,\ k_0-k_1>-1 \quad \text{(Case II)} \end{pmatrix} \tag{3.2}$$

ZERO-POWER RECTIFIER OR TOTAL-LIMITER COEFFICIENTS

Setting $f(z) = \alpha$ in (1.10),

$$\left.\begin{aligned} \frac{A_{00}}{2\alpha} &= 1 - \frac{1}{\pi^2}\int_\theta^\pi \arccos\,(k_0 + k_1\cos y)\,dy. \\ \frac{A_{10}}{\alpha} &= \frac{2}{\pi^2}\int_\theta^\pi \sqrt{1-(k_0+k_1\cos y)^2}\,dy \\ \frac{A_{01}}{\alpha} &= \frac{2k_1}{\pi^2}\int_\theta^\pi \frac{\sin^2 y\,dy}{\sqrt{1-(k_0+k_1\cos y)^2}} \\ \frac{A_{11}}{\alpha} &= \frac{2}{\pi^2}\int_\theta^\pi \cos\sqrt{1-(k_0+k_1\cos y)^2}\,dy \\ \frac{A_{20}}{\alpha} &= -\frac{2}{\pi^2}\int_\theta^\pi (k_0+k_1\cos y)\sqrt{1-(k_0+k_1\cos y)^2}\,dy \\ \frac{A_{02}}{\alpha} &= \frac{2k_1}{\pi^2}\int_\theta^\pi \frac{\sin^2 y\cos y\,dy}{\sqrt{1-(k_0+k_1\cos y)^2}} \\ \frac{A_{21}}{\alpha} &= -\frac{4}{\pi^2}\int_\theta^\pi (k_0+k_1\cos y)\cos y\sqrt{1-(k_0+k_1\cos y)^2}\,dy \end{aligned}\right\} \tag{3.3}$$

[8] E. Peterson and L. W. Hussey, Equivalent Modulator Circuits, *B. S. T. J.*, Vol. 18, pp. 32–48, Jan. 1939.

Similarly for a linear rectifier:

$$\left.\begin{aligned} \frac{A^-_{00}}{2} &= \frac{A^+_{00}}{2} + \alpha b \\ A^-_{10} &= \alpha P - A^+_{10} \\ A^-_{01} &= \alpha Q - A^+_{01} \\ A^-_{mn} &= (-)^{m+n} A^+_{mn}, \qquad m + n > 1 \end{aligned}\right\} \tag{3.4}$$

We have shown in Fig. 2 how an ideal limiting characteristic, which transmits linearly between the upper and lower limits, may be synthesized from two biased linear rectification characteristics. Equation (3.4) shows how to calculate the corresponding modulation coefficients, when the coefficients for bias of one sign are known. The limiter characteristic is equal to $\alpha z - f_1(z) - f_2(z)$, where

$$f_1(z) = \alpha \begin{pmatrix} z - b_1, & z > b_1 \\ 0, & z < b_1 \end{pmatrix} \qquad f_2(z) = \alpha \begin{pmatrix} 0, & z > -b_2 \\ z + b_2, & z < -b_2 \end{pmatrix} \tag{3.5}$$

The expression for $f_2(z)$ may also be written:

$$f_2(z) = \alpha (z + b_2) - \alpha \begin{pmatrix} z - (-b_2), & z > -b_2 \\ 0, & z < -b_2 \end{pmatrix} \tag{3.6}$$

Hence the modulation coefficient A_{mn} for the limiter may be expressed in terms of $A_{mn}(b_1)$ and $A_{mn}(-b_2)$ as follows:

$$A_{mn} = -A_{mn}(b_1) + (-)^{m+n} A_{mn}(b_2), \; m + n \neq 1 \tag{3.7}$$

If the limiter is symmetrical ($b_1 = b_2$), the even-order products vanish and the odd orders are doubled. The terms αP, αQ are to be added to the dexter of (3.7) for A_{10}, A_{01} respectively. The odd linear-rectifier coefficients, when multiplied by two, thus give the modulation products in the output of a symmetrical limiter with maximum amplitude k_0, as may be seen by substituting $b_1 = b_2 = -k_0$ in (3.7). For the fundamental components αP and αQ respectively must be subtracted from twice the A_{10} and A_{01} coefficients for k_0.

Linear Rectifier Coefficients

D.C.

$$\frac{A_{00}}{2}/\alpha P = k_0 + \frac{1}{\pi^2}\int_\theta^\pi \left[\sqrt{1 - (k_0 + k_1 \cos y)^2} - (k_0 + k_1 \cos y) \operatorname{arc}(\cos k_0 + k_1 \cos y)\right] dy \tag{3.8}$$

FUNDAMENTALS

$$A_{10}/\alpha P = 1 + \frac{1}{\pi^2}\int_\theta^\pi [(k_0 + k_1 \cos y)\sqrt{1 - (k_0 + k_1 \cos y)^2} - \arccos (k_0 + k_1 \cos y)]\, dy \tag{3.9}$$

$$A_{01}/\alpha P = k_1 + \frac{2}{\pi^2}\int_\theta^\pi [\sqrt{1 - (k_0 + k_1 \cos y)^2} - (k_0 + k_1 \cos y) \arccos (k_0 + k_1 \cos y)] \cos y\, dy \tag{3.10}$$

SUM AND DIFFERENCE PRODUCTS—Second Order

$$A_{11} = \frac{\alpha P}{\pi^2}\int_\theta^\pi [(k_0 + k_1 \cos y)\sqrt{1 - (k_0 + k_1 \cos y)^2} - \arccos (k_0 + k_1 \cos y)] \cos y\, dy \tag{3.11}$$

SUM AND DIFFERENCE PRODUCTS—Third Order

$$A_{21} = \frac{2\alpha P}{3\pi^2}\int_\theta^\pi [1 - (k_0 + k_1 \cos y)^2]^{3/2} \cos y\, dy \tag{3.12}$$

The above products are the ones usually of most interest. Others can readily be obtained either by direct integration or by use of recurrence formulas. The following set of recurrence formulas were originally derived by Mr. S. O. Rice for the biased linear rectifier:

$$\left.\begin{aligned}
&2n\, A_{mn} + k_1 (n - m - 3)\, A_{m+1,n-1} \\
&\qquad + k_1 (m + n + 3) A_{m+1,n-1} + 2k_0 n\, A_{m+1,n} = 0 \\
&2n\, A_{mn} + k_1 (n + m - 3)\, A_{m-1,n+1} \\
&\qquad + k_1 (n - m + 3) A_{m-1,n+1} + 2k_0 n\, A_{m-1,n} = 0 \\
&2m\, k_1 A_{mn} + (m - n - 3)\, A_{m-1,n+1} \\
&\qquad + (m + n + 3) A_{m+1,n+1} + 2k_0 m\, A_{m,n+1} = 0 \\
&2\, m\, k_1 A_{mn} + (m + n - 3)\, A_{m-1,n-1} \\
&\qquad + (m - n + 3) A_{m+1,n-1} + 2k_0 m\, A_{m,n-1} = 0
\end{aligned}\right\} \tag{3.13}$$

By means of these relations, all products can be expressed in terms of A_{00}, A_{10}, A_{01}, and A_{11}. The following specific results are tabulated:

$$\left.\begin{aligned}
A_{20} &= \tfrac{1}{3}(A_{00} - 2k_1 A_{11} - 2k_0 A_{10}) \\
A_{02} &= \frac{1}{3k_1}(k_1 A_{00} - 2A_{11} - 2k_0 A_{01})
\end{aligned}\right\} \tag{3.14}$$

$$\left.\begin{aligned} A_{21} &= \tfrac{1}{2}(A_{01} - k_1 A_{10} - k_0 A_{11}) \\ A_{12} &= \frac{1}{2k_1}(k_1 A_{10} - A_{01} - k_0 A_{11}) \end{aligned}\right\} \tag{3.15}$$

$$\left.\begin{aligned} A_{30} &= -k_0 A_{20} - k_1 A_{21} \\ A_{03} &= -\frac{1}{k_1}(k_0 A_{02} + A_{12}) \end{aligned}\right\} \tag{3.16}$$

The third-order product A_{21} is of considerable importance in the design of carrier amplifiers and radio transmitters, since the $(2p - q)$-product is the cross-product of lowest order falling back in the fundamental band when overload occurs. Figure 14 shows curves of A_{21} calculated by Mr. J. O. Edson from Eq. (3.12) by mechanical integration.

We point out also that the linear-rectifier coefficients give the Fourier series expansion of the admittance of a biased square-law rectifier when two frequencies are applied.

We shall next discuss the problem of reduction of the integrals appearing above to a closed form in terms of tabulated elliptic integrals[9]. This can be done for all the coefficients above except the d-c for the zero-power law and for the d-c and two fundamentals for the linear rectifier. These contain the integral

$$\Xi(k_0, k_1) = \int_\theta^\pi \arccos(k_0 + k_1 \cos y)\, dy \tag{3.17}$$

which has been calculated separately and plotted in Fig. 22. When the arc cos term is accompanied by $\cos my$ as a multiplier with $m \neq 0$, an integration by parts is sufficient to reduce the integrand to a rational function of $\cos y$ and the radical $\sqrt{1 - (k_0 + k_1 \cos y)^2}$, which may be reduced at once to a recognizable elliptic integral by the substitution $z = \cos y$. It is found that all the integrals except that of (3.17) appearing in the results can be expressed as the sum of a finite number of integrals of the form:

$$Z_m = \int_{-1}^{\cos\theta} \frac{z^m\, dz}{\sqrt{(1 - z^2)[1 - (k_0 + k_1 z)^2]}}, \quad m = 0, 1, 2, \cdots \tag{3.18}$$

By differentiating the expression $z^{m-3}\sqrt{(1 - z)^2[1 - (k_0 + k_1 z^2]}$ with respect to z, we may derive the recurrence formula:

$$\begin{aligned} Z_m = -\frac{1}{(m-1)k_1^2}[&(2m - 3)k_0 k_1 z_{m-1} \\ &+ (m - 2)(k_0^2 - k_1^2 - 1)Z_{m-2} \\ &- (2m - 5)k_0 k_1 Z_{m-3} + (m - 3)(1 - k_0^2)Z_{m-4}] \end{aligned} \tag{3.19}$$

[9] Power series expansions of coefficients such as treated here have been given by A. G. Tynan, Modulation Products in a Power Law Modulator, *Proc. I. R. E.*, Vol. 21, pp. 1203–1209, Aug. 1933.

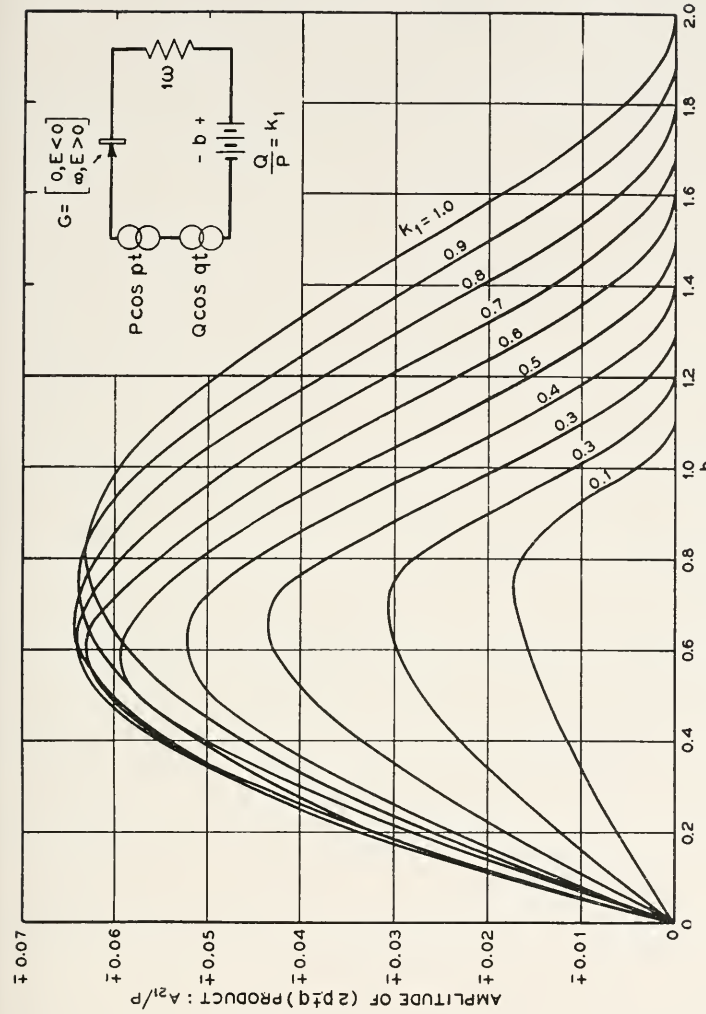


Fig. 14.—$(2p \pm q)$—Product from linear rectifier with bias.

It thus is found that the value of Z_m for all values of m greater than 2 can be expressed in terms of Z_0, Z_1, and Z_2.

Eq. (3.18) may be written in the form:

$$Z_m = \frac{1}{k_1}\int_{z_2}^{z_3} \frac{z^m\,dz}{\sqrt{(z-z_1)\,(z-z_2)(z_3-z)(z_4-z)}} \tag{3.20}$$

$$\left.\begin{aligned} z_1 &= -(1+k_0)/k_1\,,\ z_2 = -1 \\ z_3 &= \begin{pmatrix} (1-k_0)/k_1\,,\ \text{Case I} \\ 1,\ \text{Case II} \end{pmatrix} \\ z_4 &= \begin{pmatrix} 1,\ \text{Case I} \\ (1-k_0)/k_1,\ \text{Case II}, \end{pmatrix} \end{aligned}\right\} \tag{3.21}$$

The substitution

$$z = \frac{z_2(z_3-z_1) - z_1(z_3-z_2)u^2}{z_3 - z_1 - (z_3-z_2)u^2} \tag{3.22}$$

reduces the integral to

$$Z_m = \frac{2}{k_1\sqrt{(z_4-z_2)(z_3-z_1)}} \int_0^1 \frac{\left(z_1 + \dfrac{z_2-z_1}{1-\eta\,u^2}\right)^m du}{\sqrt{(1-u^2)(1-\kappa^2u^2)}} \tag{3.23}$$

where:

$$\eta = \frac{z_3-z_2}{z_3-z_1} \tag{3.24}$$

$$x^2 = \frac{(z_4-z_1)(z_3-z_2)}{(z_4-z_2)(z_3-z_1)} \tag{3.25}$$

Hence if K, E and Π represent respectively complete elliptic integrals of the first, second, and third kinds with modulus κ, and in the case of third kind with parameter $-\eta$, we have immediately:

$$Z_0 = \frac{2K}{k_1\sqrt{(z_4-z_2)(z_3-z_1)}} \tag{3.26}$$

$$Z_1 = \frac{2[z_1K + (z_2-z_1)\,\Pi]}{k_1\sqrt{(z_4-z_2)(z_3-z_1)}} \tag{3.27}$$

$$\begin{aligned} Z_2 = \frac{2}{k_1\sqrt{(z_4-z_2)(z_3-z_1)}} \Bigg[& z_1^2K + 2z_1(z_2-z_1)\Pi \\ & + (z_2-z_1)^2 \int_0^1 \frac{du}{(1-\eta\,u^2)^2\sqrt{(1-u^2)\,(1-\kappa^2u^2)}} \Bigg] \end{aligned} \tag{3.28}$$

To complete the evaluation of Z_2, assume a relation of the following type with undetermined constants C_1, C_2, C_3, C_4:

$$\int_0^z \frac{du}{(1-\eta u^2)^2\sqrt{(1-u^2)(1-\kappa^2 u^2)}} = C_1\int_0^z \frac{du}{\sqrt{(1-u^2)(1-\kappa^2 u^2)}} + C_2\int_0^z \sqrt{\frac{1-\kappa^2 u^2}{1-u^2}}\,du + C_3\int_0^z \frac{du}{(1-\eta u^2)\sqrt{(1-u^2)(1-\kappa^2 u^2)}} + C_4\frac{z\sqrt{(1-z^2)(1-\kappa^2 z^2)}}{1-\eta z^2} \tag{3.29}$$

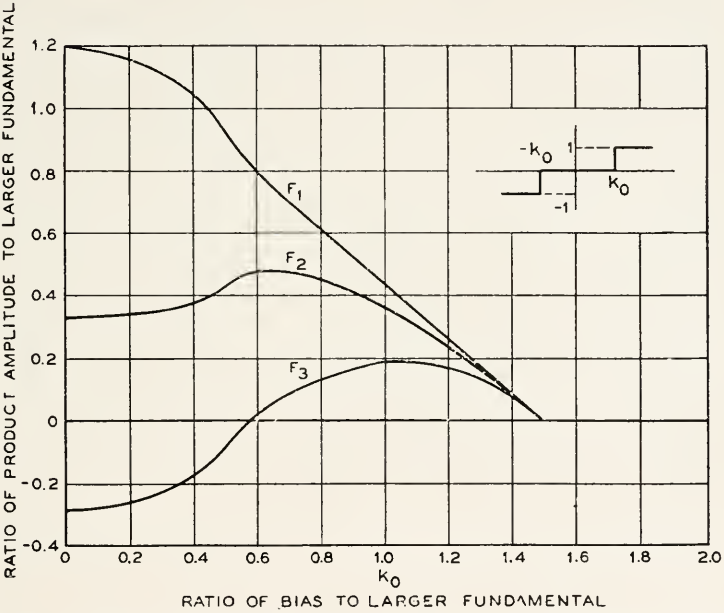


Fig. 15.—Fundamentals and $(2p \pm q)$—product from full-wave biased zero-power-law rectifier with ratio of applied fundamental amplitudes equal to 0.5. F_1 = larger fundamental, F_2 = smaller fundamental, $F_3 = (2p \pm q)$—product.

Differentiate both sides with respect to z, set $z = 1$, and clear fractions. Equating coefficients of like powers of z separately then gives four simultaneous equations in C_1, C_2, C_3, C_4. Solving for C_1, C_2, C_3 and setting $z = 1$ in (3.29) gives

$$\int_0^1 \frac{du}{(1-\eta u^2)^2\sqrt{(1-u^2)(1-\kappa^2 u^2)}} = \frac{1}{2(\eta-1)}\left[K + \frac{\eta E}{\kappa^2-\eta} + \frac{(2\eta-3)\kappa^2 - \eta(\eta-2)}{\kappa^2-\eta}\,\Pi\right] \tag{3.30}$$

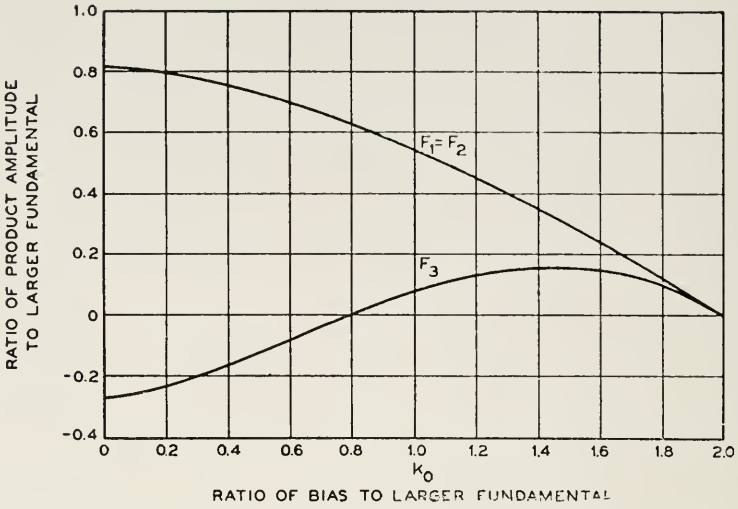


Fig. 16.—Fundamentals and $(2p \pm q)$—product from full-wave biased zero-power-law rectifier with equal applied fundamental amplitudes.

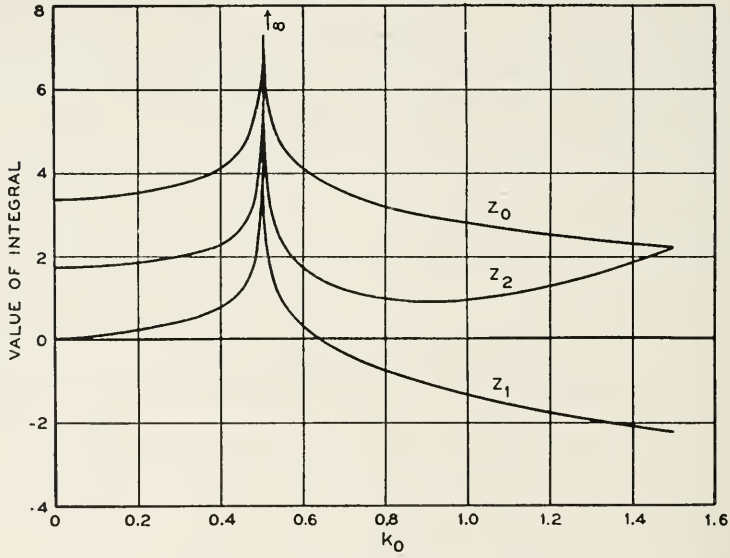


Fig. 17.—The integral Z_m with $k_1 = 0.5$.

Since the necessary tables of II are not available, we make use of Legendre's Transformation,[10] which in this case gives:

[10] Legendre, Traités des Fonctions Elliptiques, Paris, 1825–28, Vol. I, Ch. XXIII.

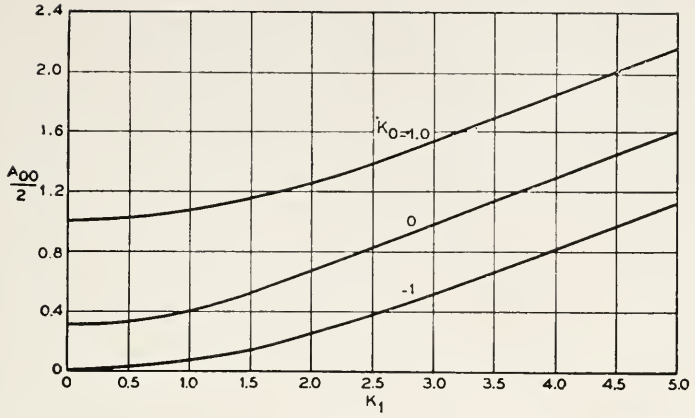


Fig. 18.—D-c. term in linear rectifier output with two applied frequencies.

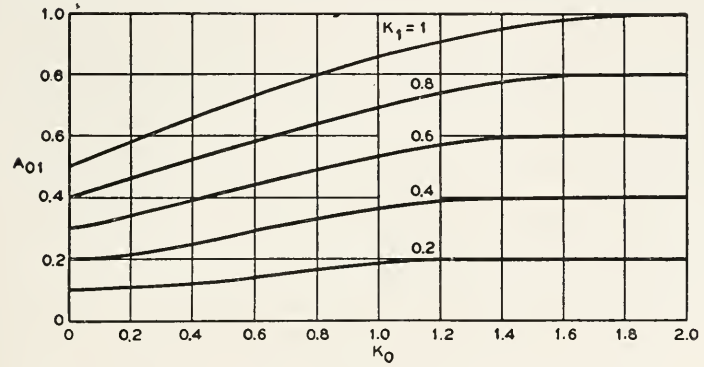


Fig. 19.—Smaller fundamental in biased linear rectifier output.

$$\Pi = K + \frac{\tan \phi}{\sqrt{1 - \kappa^2 \sin^2 \phi}} [KE(\phi) - EF(\phi)] \tag{3.31}$$

$$\phi = \arcsin \frac{\eta^{1/2}}{\kappa} \tag{3.32}$$

$$F(\phi) = \int_0^{\phi} \frac{d\theta}{\sqrt{1 - \kappa^2 \sin^2 \theta}} \tag{3.33}$$

$$E(\phi) = \int_0^{\phi} \sqrt{1 - \kappa^2 \sin^2 \theta}\, d\theta \tag{3.34}$$

The functions $F(\phi)$ and $E(\phi)$ are incomplete elliptic integrals of the first and second kinds. They are tabulated in a number of places. Fairly good tables, e.g. the original ones of Legendre, are needed here since the difference between $KE(\phi)$ and $EF(\phi)$ is relatively small.

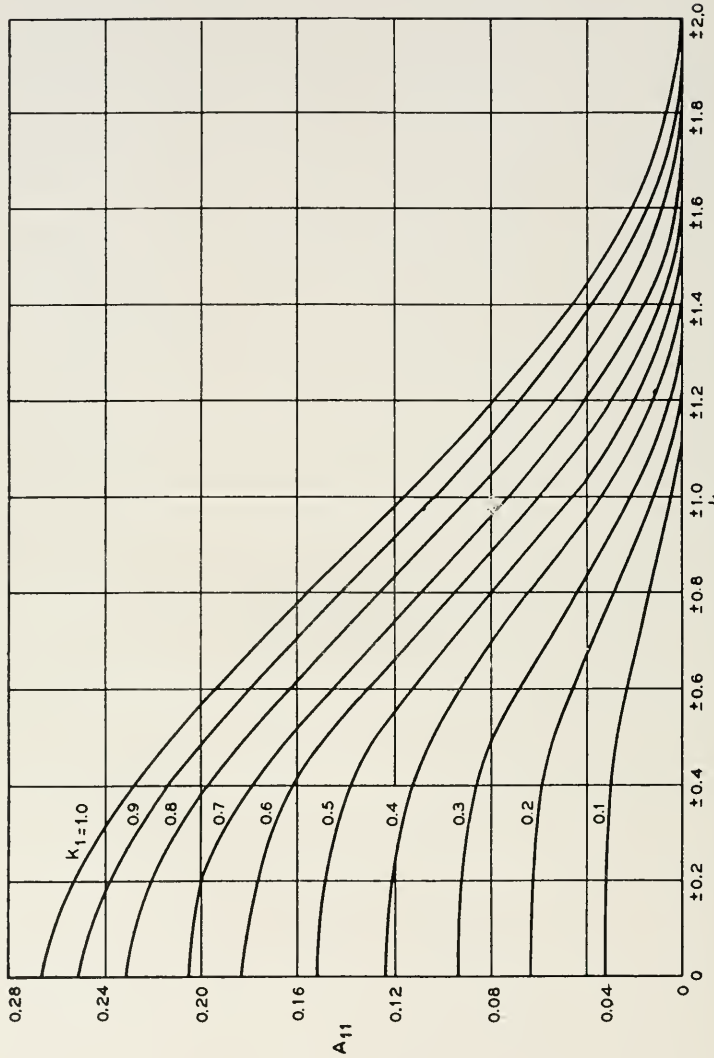


Fig. 20.—Second-order sideband amplitude in biased linear rectifier output.

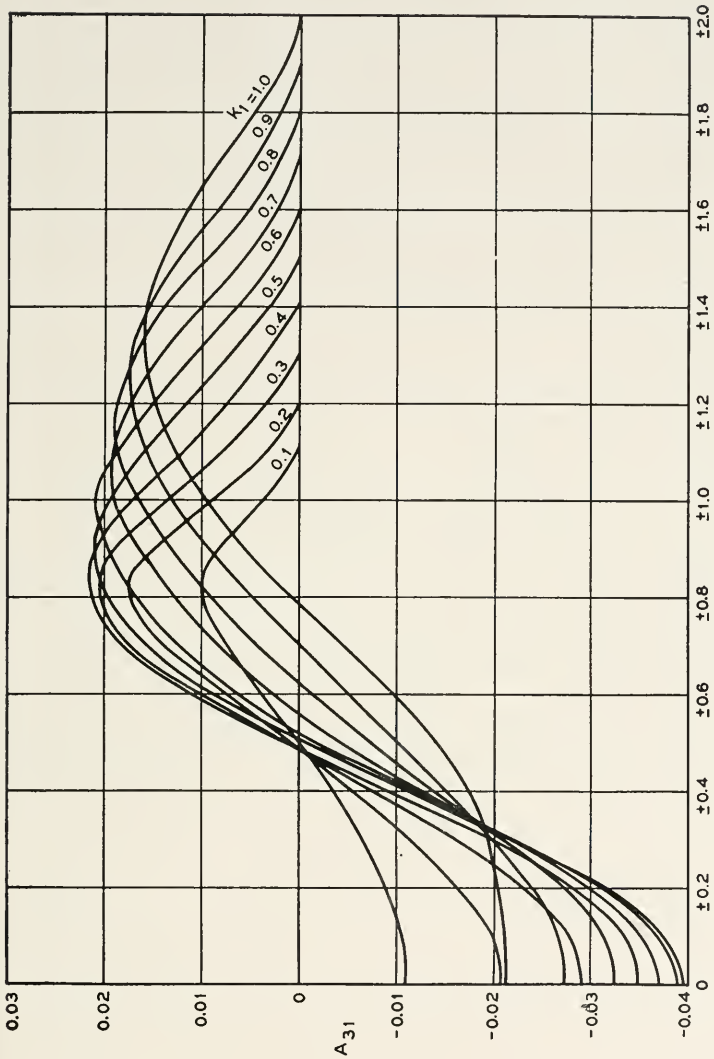


Fig. 21.—$(3p \pm q)$—product in biased linear rectifier output.

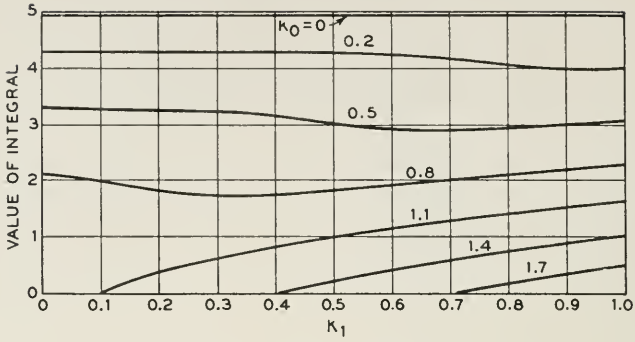


Fig. 22.—Graph of the integral $\Xi\ (k_0,\ k_1)$.

Summarizing:

$$\left.\begin{aligned}
&\text{Case I, } k_0 + k_1 > 1, \qquad k_0 - k_1 < 1 \\
&Z_0 = \frac{K}{\sqrt{k_1}} \\
&Z_1 = \frac{2}{k_1}\left[KE(\phi) - EF(\phi)\right] - Z_0 \\
&Z_2 = \frac{1}{k_1}\left[Z_0 - k_0 Z_1 - \frac{2}{\sqrt{k_1}} E\right] \\
&\kappa = \sqrt{\frac{(k_1+1)^2 - k_0^2}{4k_1}} \\
&\phi = \arcsin \sqrt{\frac{2k_1}{1 + k_0 + k_1}}
\end{aligned}\right\} \tag{3.35}$$

Case II, $k_0 + k_1 < 1,\ k_0 - k_1 > -1$

$$\left.\begin{aligned}
&Z_0 = \frac{2K}{\sqrt{(1+k_1)^2 - k_0^2}} \\
&Z_1 = \frac{2}{k_1}\left[KE(\phi) - EF(\phi)\right] - Z_0 \\
&Z_2 = \frac{1}{2k_1^2}\left[(1 + k_1^2 - k_0^2)Z_0 - 2k_0 k_1 Z_1 - 2E\sqrt{(1+k_1)^2 - k_0^2}\right] \\
&\kappa = \sqrt{\frac{4k_1}{(1+k_1)^2 - k_0^2}} \\
&\phi = \arcsin \sqrt{\frac{1 - k_0 + k_1}{2}}
\end{aligned}\right\} \tag{3.36}$$

The values of the fundamentals and third-order sum and difference products for the biased zero-power-law rectifier have been calculated by the formulas above for the cases $k_1 = .5$ and $k_1 = 1$. The resulting curves are shown in Fig. (15) and (16). The values of the auxiliary integrals Z_0, Z_1, and Z_2 are shown for $k_1 = .5$ in Fig. (17). These integrals become infinite at $k_0 = 1 - k_1$ so that the formulas for the modulation coefficients become indeterminate at this point. The limiting values can be evaluated from the integrals (3.3), etc., directly in terms of elementary functions when the relation $k_0 = 1 - k_1$ is substituted, except for the Ξ-function.

Limiting forms of the coefficients when k_0 is small are of value in calculating the effect of a small signal superimposed on the two sinusoidal components in an unbiased rectifier. By straightforward power-series expansion in k_0, we find:

Zero-Power-Law Rectifier, k_0 Small:

$$\left.\begin{aligned} A_{10} &= \frac{4}{\pi^2}E - \frac{2E}{\pi^2(1-k^2)}k_0^2 + \cdots \\ A_{01} &= \frac{4}{\pi^2 k_1}[E - (1-k_1^2)K] + \frac{2}{\pi^2 k_1}\left(\frac{E}{1-k_1^2} - K\right)k_0^2 + \cdots \\ A_{21} &= \frac{4}{3\pi^2 k_1}[(1-2k_1^2)E - (1-k_1^2)K] \\ &\quad + \frac{2}{\pi^2 k_1}\left[K - \frac{1-2k_1^2}{1-k_1^2}E\right]k_0^2 + \cdots \end{aligned}\right\} \quad (3.37)$$

In the above expressions, the modulus of K and E is k_1. When $k_0 = 0$, these coefficients reduce to half the values of the full-wave unbiased zero-power-law coefficients, which have been tabulated in a previous publication.[11]

Acknowledgment

In addition to the persons already mentioned, the writer wishes to thank Miss M. C. Packer, Miss J. Lever and Mrs. A. J. Shanklin for their assistance in the calculations of this paper.

[11] R. M. Kalb and W. R. Bennett, Ferromagnetic Distortion of a Two-Frequency Wave, *B. S. T. J.*, Vol. XIV, April 1935, Eq. (21), p. 336.

Properties and Uses of Thermistors—Thermally Sensitive Resistors*

By J. A. BECKER, C. B. GREEN and G. L. PEARSON

A new circuit element and control device, the thermistor or thermally sensitive resistor, is made of solid semiconducting materials whose resistance decreases about four per cent per centigrade degree. The thermistor presents interesting opportunities to the designer and engineer in many fields of technology for accomplishing tasks more simply, economically and better than with available devices. Part I discusses the conduction mechanism in semiconductors and the criteria for usefulness of circuit elements made from them. The fundamental physical properties of thermistors, their construction, their static and dynamic characteristics and general principles of operation are treated.

Part II of this paper deals with the applications of thermistors. These include: sensitive thermometers and temperature control elements, simple temperature compensators, ultrahigh frequency power meters, automatic gain controls for transmission systems such as the Types K2 and L1 carrier telephone systems, voltage regulators, speech volume limiters, compressors and expandors, gas pressure gauges and flowmeters, meters for thermal conductivity determination of liquids, and contactless time delay devices. Thermistors with short time constants have been used as sensitive bolometers and show promise as simple compact audio-frequency oscillators, modulators and amplifiers.

PART I—PROPERTIES OF THERMISTORS

Introduction

THERMISTORS, or *therm*ally sensitive re*sistors*, are devices made of solids whose electrical resistance varies rapidly with temperature. Even though they are only about 15 years old they have already found important and large scale uses in the telephone plant and in military equipments. Some of these uses are as time delay devices, protective devices, voltage regulators, regulators in carrier systems, speech volume limiters, test equipment for ultra-high-frequency power, and detecting elements for very small radiant power. In all these applications thermistors were chosen because they are simple, small, rugged, have a long life, and require little maintenance. Because of these and other desirable properties, thermistors promise to become new circuit elements which will be used extensively in the fields of communications, radio, electrical and thermal instrumentation, research in physics, chemistry and biology, and war technology. Specific types of uses which will be discussed in the second part of this paper include: 1) simple, sensitive and fast responding thermometers,

* Published in *Elec. Engg.*, November 1946.

The authors acknowledge their indebtedness to Messrs. J. H. Scaff and H. C. Theuerer for furnishing samples for most of the curves in Fig. 4, and to Mr. G. K. Teal for the data for the lowest curve in that figure.

temperature compensators and temperature control devices; 2) special switching devices without moving contacts; 3) regulators or volume limiters; 4) pressure gauges, flowmeters, and simple meters for measuring thermal conductivity in liquids and gases; 5) time delay and surge suppressors; 6) special oscillators, modulators and amplifiers for relatively low frequencies. Before these uses are discussed in detail it is desirable to present the physical principles which determine the properties of thermistors.

The question naturally arises "why have devices of this kind come into use only recently?" The answer is that thermistors are made of semiconductors and that the resistance of these can vary by factors up to a thousand or a million with surprisingly small amounts of certain impurities, with heat treatment, methods of making contact and with the treatment during life or use. Consequently the potential application of semiconductors was discouraged by experiences such as the following: two or more units made by what appeared to be the same process would show large variations in their properties. Even the same unit might change its resistance by factors of two to ten by exposure to moderate temperatures or to the passage of current. Before semiconductors could seriously be considered in industrial applications, it was necessary to devote a large amount of research and development effort to a study of the nature of the conductivity in semiconductors, and of the effect of impurities and heat treatment on this conductivity, and to methods of making reliable and permanent contacts to semiconductors. Even though Faraday discovered that the resistance of silver sulphide changed rapidly with temperature, and even though thousands of other semiconductors have been found to have large negative temperature coefficients of resistance, it has taken about a century of effort in physics and chemistry to give the engineering profession this new tool which may have an influence similar to that of the vacuum tube and may replace vacuum tubes in many instances.

If thermistors are to be generally useful in industry:

1) it should be possible to reproduce units having the same characteristics;
2) it should be possible to maintain constant characteristics during use; the contact should be permanent and the unit should be chemically inert;
3) the units should be mechanically rugged;
4) the technique should be such that the material can be formed into various shapes and sizes;
5) it should be possible to cover a wide range of resistance, temperature coefficient and power dissipation.

Thermistors might be made by any method by which a semiconductor

could be shaped to definite dimensions and contacts applied. These methods include: 1) melting the semiconductor, cooling and solidifying, cutting to size and shape; 2) evaporation; 3) heating compressed powders of semiconductors to a temperature at which they sinter into a strong compact mass and firing on metal powder contacts. While all three processes have been used, the third method has been found to be most generally useful for mass production. This method is similar to that employed in ceramics or in powder metallurgy. At the sintering temperatures the powders recrystallize and the dimensions shrink by controlled amounts. The powder process makes it possible to mix two or more semiconducting oxides in varying proportions and obtain a homogeneous and uniform solid. It is thus possible to cover a considerable range of specific resistance and tem-

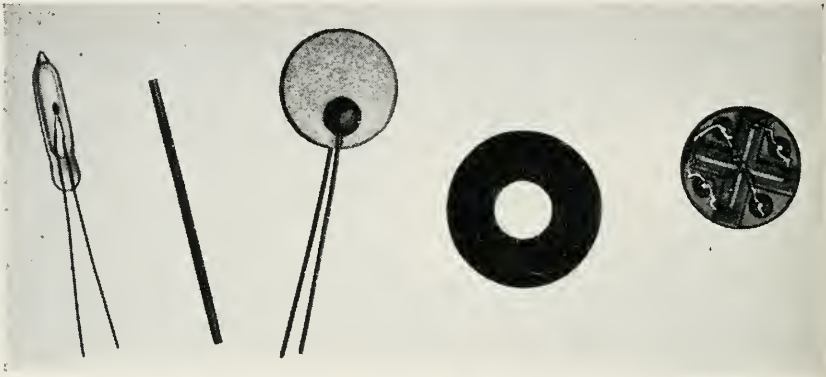


Fig. 1.—Thermistors made in the form of a bead, rod, disc, washer and flakes.

perature coefficient of resistance with the same system of oxides. By means of the powder process it is possible to make thermistors of a great variety of shapes and sizes to cover a large range of resistances and power handling capacities.

Figure 1 is a photograph of thermistors made in the form of beads, rods, discs, washers and flakes. Beads are made by stringing two platinum alloy wires parallel to each other with a spacing of five to ten times the wire diameter. A mass of a slurry of mixed oxides is applied to the wires. Surface tension draws this mass into the form of a bead. From 10 to 20 such beads are evenly spaced along the wires. The beads are allowed to dry and are heated slightly until they have sufficient strength so that the string can be handled. They then are passed through the sintering furnace. The oxides shrink onto the platinum alloy wires and make an intimate and permanent electrical contact. The wires then are cut to separate the individual beads.

The diameters of the beads range from 0.015 to 0.15 centimeters with wire diameters ranging from 0.0025 to 0.015 centimeters.

Rod thermistors are made by mixing the oxides with an organic binder and solvent, extruding the mixture through a die, drying, cutting to length, heating to drive out the binder, and sintering at a high temperature. Contacts are applied by coating the ends with silver, gold, or platinum paste as used in the ceramic art, and heating or curing the paste at a suitable temperature. The diameter of the rods can ordinarily be varied from 0.080 to 0.64 centimeter. The length can vary from 0.15 to 5 centimeters.

Discs and washers are made in a similar way by pressing the bonded powders in a die. Possible disc diameters are 0.15 to 3 or 5 centimeters; thicknesses from 0.080 to 0.64 centimeter.

Flakes are made by mixing the oxides with a suitable binder and solvent to a creamy consistency, spreading a film on a smooth glass surface, allowing the film to dry, removing the film, cutting it into flakes of the desired size and shape, and firing the flakes at the sintering temperatures on smooth ceramic surfaces. Contacts are applied as described above. Possible dimensions are: thickness, 0.001 to 0.004 centimeter; length, 0.1 to 1.0 centimeter; width, 0.02 to 0.1 centimeter.

In any of these forms lead wires can be attached to the contacts by soldering or by firing heavy metal pastes. The dimensional limits given above are those which have been found to be readily attainable.

In the design of a thermistor for a specific application, the following characteristics should be considered: 1) Mechanical dimensions including those of the supports. 2) The material from which it is made and its properties. These include the specific resistance and how it varies with temperature, the specific heat, density, and expansion coefficient. 3) The dissipation constant and power sensitivity. The dissipation constant is the watts that are dissipated in the thermistor divided by its temperature rise in centigrade degrees above its surroundings. The power sensitivity is the watts dissipated to reduce the resistance by one per cent. These constants are determined by the area and nature of the surface, the surrounding medium, and the thermal conductivity of the supports. 4) The heat capacity which is determined by specific heat, dimensions, and density. 5) The time constant. This determines how rapidly the thermistor will heat or cool. If a thermistor is heated above its surroundings and then allowed to cool, its temperature will decrease rapidly at first and then more slowly until it finally reaches ambient temperature. The time constant is the time required for the temperature to fall 63 per cent of the way toward ambient temperature. The time constant in seconds is equal to the heat capacity in joules per centigrade degree divided by the dissipation constant in watts

per centigrade degree. 6) The maximum permissible power that can be dissipated consistent with good stability and long life, for continuous operation, and for surges. This can be computed from the dissipation constant and the maximum permissible temperature rise. This and the resistance-temperature relation determine the maximum decrease in resistance.

Properties of Semiconductors

As most thermistors are made of semiconductors it is important to discuss the properties of the latter. A semiconductor may be defined as a substance

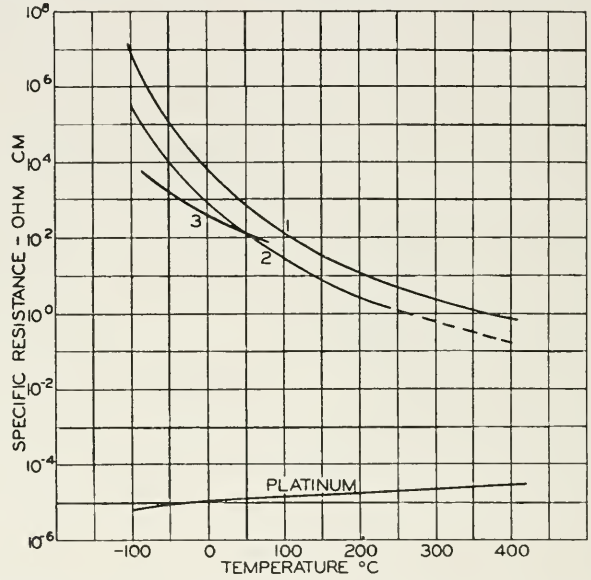


Fig. 2.—Logarithm of specific resistance versus temperature for three thermistor materials as compared with platinum.

whose electrical conductivity at or near room temperature is much less than that of typical metals but much greater than that of typical insulators. While no sharp boundaries exist between these classes of conductors, one might say that semiconductors have specific resistances at room temperature from 0.1 to 10^9 ohm centimeters. Semiconductors usually have high negative temperature coefficients of resistance. As the temperature is increased from 0°C. to 300°C., the resistance may decrease by a factor of a thousand. Over this same temperature range the resistance of a typical metal such as platinum will increase by a factor of two. Figure 2 shows how the logarithm of the specific resistance, ρ, varies with temperature, T, in degrees centigrade for three typical semiconductors and for platinum.

Curves 1 and 2 are for Materials No. 1 and No. 2 which have been extensively used to date. Material No. 1 is composed of manganese and nickel oxides. Material No. 2 is composed of oxides of manganese, nickel and cobalt. The dashed part of Curve 2 covers a region in which the resistance-temperature relation is not known as accurately as it is at lower temperatures. Curve 3 is an experimental curve for a mixture of iron and zinc

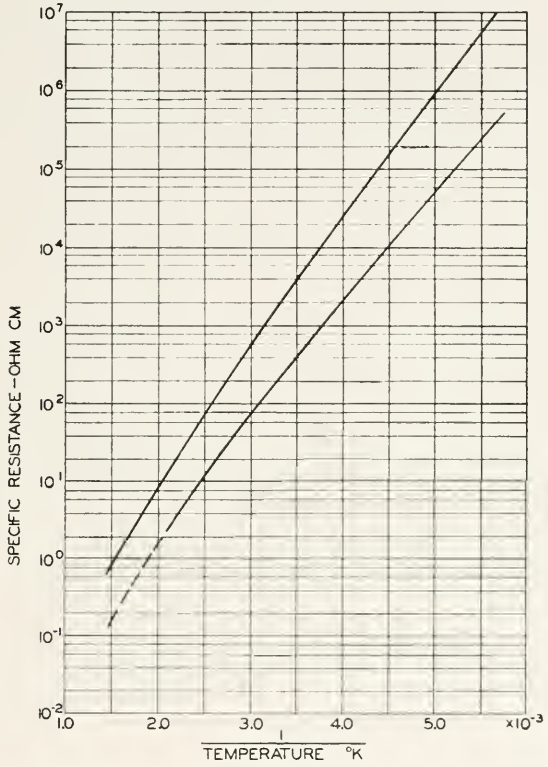


Fig. 3.—Logarithm of the specific resistance of two thermistor materials as a function of inverse absolute temperature. See equation (1).

oxides in the proportions to form zinc ferrite. From Fig. 2 it is obvious that neither the resistance R nor log R varies linearly with T.

Figure 3 shows plots of log ρ versus $1/T$ for Materials No. 1 and No. 2. These do form approximate straight lines. Hence

$$\rho = \rho_\infty \epsilon^{B/T} \text{ or } \rho = \rho_0 \epsilon^{(B/T)-(B/T_0)} \qquad (1)$$

where T = temperature in degrees Kelvin; $\rho_\infty = \rho$ when $T = \infty$ or $1/T = 0$; $\rho_0 = \rho$ when $T = T_0$; ϵ = Naperian base = 2.718 and B is a constant equal to 2.303 times the slope of the straight lines in Fig. 3. The dimensions of B

are Kelvin degrees or centigrade degrees; it plays the same role in equation (1) as does the work function in Richardson's equation for thermionic emission. For Material No. 1, $B = 3920C°$. This corresponds to an electron energy equivalent to 3920/11600 or 0.34 volt.

While the curves in Fig. 3 are approximately straight, a more careful investigation shows that the slope increases linearly as the temperature increases. From this it follows that a more precise expression for ρ is:

$$\rho = AT^{-c}\,\epsilon^{D/T} \qquad \text{or}$$

$$\log \rho = \log A - c \log T + D/2.303T \tag{2}$$

The constant c is a small positive or negative number or zero. For Material No. 1, $\log A = 5.563$, $c = 2.73$ and $D = 3100$. For a particular form of Material No. 2 $\log A = 11.514$, $c = 4.83$ and $D = 2064$.

If we define temperature coefficient of resistance, α, by the equation

$$\alpha = (1/R)\,(dR/dT) \tag{3}$$

it follows from equation (1) that

$$\alpha = -B/T^2. \tag{4}$$

For Material No. 1 and $T = 300°K$, $\alpha = -3920/90{,}000 = -0.044$. For platinum, $\alpha = +0.0037$ or roughly ten times smaller than for semiconductors and of the opposite sign. From equation (2) it follows that

$$\alpha = -(D/T^2) - (c/T). \tag{5}$$

From equation (3) it follows that

$$\alpha = (1/2.303)\,(d \log R/dT). \tag{6}$$

For a discussion of the nature of the conductivity in semiconductors, it is simpler and more convenient to consider the conductivity, σ, rather than the resistivity, ρ.

$$\sigma = 1/\rho \qquad \text{and} \qquad \log \sigma = -\log \rho. \tag{7}$$

The characteristics of semiconductors are brought out more clearly if the conductivity or its logarithm are plotted as a function of $1/T$ over a wide temperature range. Figure 4 is such a plot for a number of silicon samples containing increasing amounts of impurity. At high temperatures all the samples have nearly the same conductivity. This is called the intrinsic conductivity since it seems to be an intrinsic property of silicon. At low temperatures the conductivity of different samples varies by large factors. In this region silicon is said to be an impurity semiconductor. For extremely pure silicon only intrinsic conductivity is present and the

resistivity obeys equation (1). As the concentration of a particular impurity increases, the conductivity increases and the impurity conductivity predominates to higher temperatures. Some impurities are much more effective in increasing the conductivity than others. One hundred parts per million of some impurities may increase the conductivity of pure silicon at room temperature by a factor of 10^7. Other impurities may be present

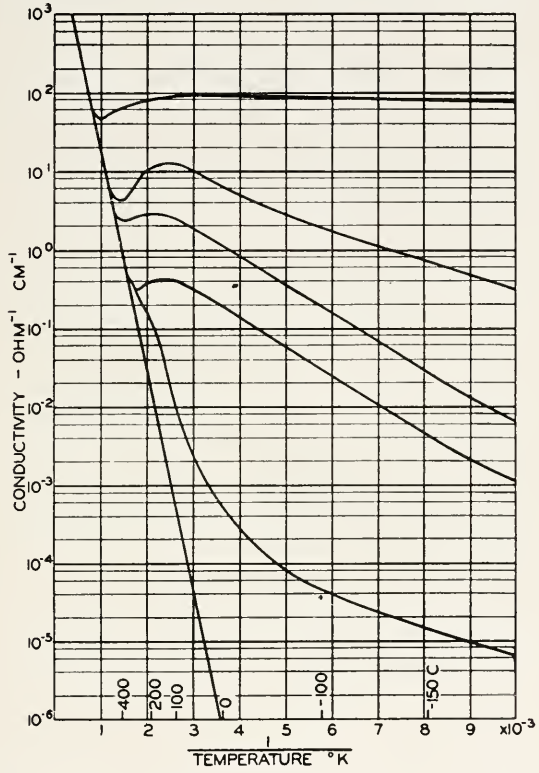


Fig. 4.—Logarithm of the conductivity of various specimens of silicon as a function of inverse absolute temperature. The conductivity increases with the amount of impurity.

in 10,000 parts per million and have a small effect on the conductivity. Two samples may contain the same concentration of an impurity and still differ greatly in their low temperature conductivity; if the impurity is in solid solution, i.e., atomically dispersed, the effect is great; if the impurity is segregated in atomically large particles, the effect is small. Since heat treatments affect the dispersion of impurities in solids, the conductivity of semiconductors may frequently be altered radically by heat treatment. Some other semiconductors are not greatly affected by heat treatment.

The impurity need not even be a foreign element; in the case of oxides or sulphides, it can be an excess or a deficiency of oxygen or sulphur from the exact stoichiometric relation. This excess or deficiency can be brought about by heat treatment. Figure 5 shows how the conductivity depends on temperature for a number of samples of cuprous oxide, Cu_2O, heat

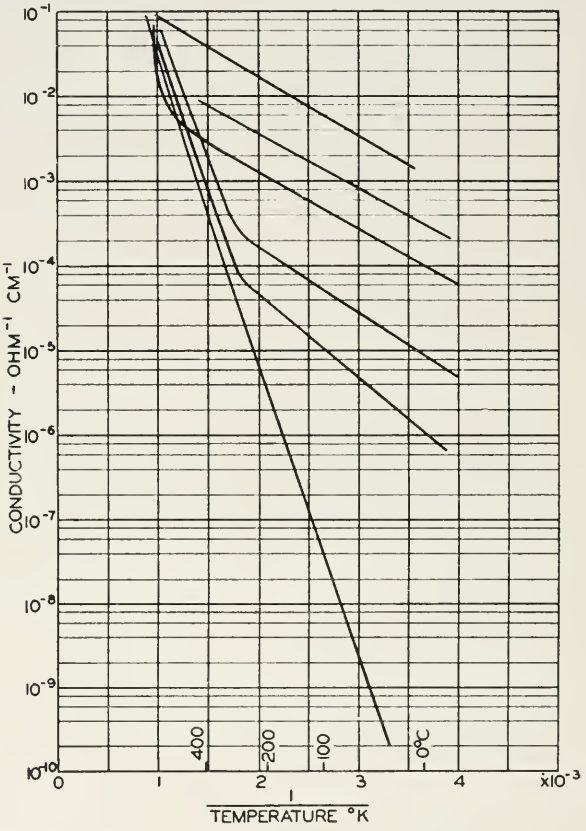


Fig. 5.—Logarithm of the conductivity of various specimens of cuprous oxide as a function of inverse absolute temperature. The conductivity increases with the amount of excess oxygen above the stoichiometric value in Cu_2O. Data from reference 1.

treated in such a way as to result in varying amounts of excess oxygen from zero to about one per cent.[1] The greater the amount of excess oxygen the greater is the conductivity in the low temperature range. At high temperatures, all samples have about the same conductivity.

Semiconductors can be classified on the basis of the carriers of the current into ionic, electronic, and mixed conductors. Chlorides such as NaCl and some sulphides are ionic semiconductors; other sulphides and a few oxides

such as uranium oxide are mixed semiconductors; electronic semiconductors include most oxides such as Mn_2O_3, Fe_2O_3, NiO, carbides such as silicon carbide, and elements such as boron, silicon, germanium and tellurium. In ionic and mixed conductors, ions are transported through the solid. This changes the density of carriers in various regions, and thus changes the conductivity. Because this is undesirable, they are rarely used in making thermistors, and hence we will concentrate our interest on electronic semiconductors.

The theoretical and experimental physicists have established that there are two types of electronic semiconductors which can be called *N* and *P* type, depending upon whether the carriers are negative electrons or are equivalent to positive "holes" in the filled energy band. In *N* type, the

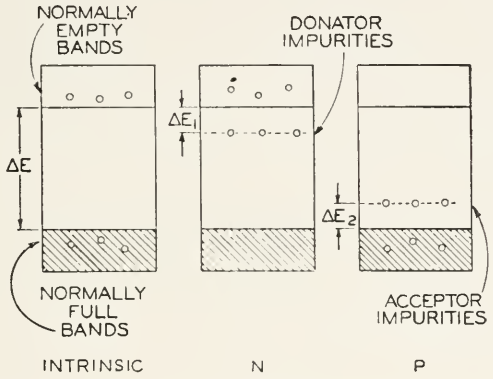


Fig. 6.—Schematic energy level diagrams illustrating intrinsic, *N* and *P* types of semiconductors.

carriers are deflected by a magnetic field as negatively charged particles would be and conversely for *P* type. The direction of deflections is ascertained by measurement of the sign of the Hall effect. The direction of the thermoelectric effect also fixes the sign of the carriers. By determining the resistivity, Hall coefficient and thermoelectric power of a particular specimen at a particular temperature it is possible to determine the density of carriers, whether they are negative or positive, and their mobility or mean free path. The mobility is the mean drift velocity in a field of one volt per centimeter.

The existence of these classifications is explained by the theoretical physicist[2,3,4] in terms of the diagrams in Fig. 6. In an intrinsic semiconductor at low temperatures the valence electrons completely fill all the allowable energy states. According to the exclusion principle only one electron can occupy a particular energy state in any system. In semiconductors and

insulators there exists a region of energy values, just above the allowed band, which are not allowed. The height of this unallowed band is expressed in equivalent electron volts, ΔE. Above this unallowed band there exists an allowed band; but at low temperatures there are no electrons in this band. When a field is applied across such a semiconductor, no electron can be accelerated, because if it were accelerated its energy would be increased to an energy state which is either filled or unallowed. As the temperature is raised some electrons acquire sufficient energy to be raised across the unallowed band into the upper allowed band. These electrons can be accelerated into a slightly higher energy state by the applied field and thus can carry current. For every electron that is put into an "activated" state there is left behind a "hole" in the normally filled band. Other electrons having slightly lower energies can be accelerated into these holes by the applied field. The physicist has shown that these holes act toward the applied field as if they were particles having a charge equal to that of an electron but of opposite sign and a mass equal to or somewhat larger than the electronic mass. In an intrinsic semiconductor about half the conductivity is due to electrons and half due to holes.

The quantity ΔE is related to B in equation (1) by:

$$2B = (\Delta E)\, e/k \tag{8}$$

in which B is in centigrade degrees, ΔE is in volts, e is the electronic charge in coulombs, k is Boltzmann's constant in joules per centigrade degree. The value of e/k is 11,600 so that

$$\Delta E = B/5800. \tag{8a}$$

The difference between metals, semiconductors, and insulators results from the value of ΔE. For metals ΔE is zero or very small. For semiconductors ΔE is greater than about 0.1 volt but less than about 1.5 volts. For insulators ΔE is greater than about 1.5 volts.

Some impurities with positive valencies which may be present in the semiconductor may have energy states such that ΔE_1 volts equivalent energy can raise the valence electron of the impurity atom into the allowed conduction band. See Figure 6. The electron now can take part in conduction; the donator impurity is a positive ion which is usually bound to a particular location and can take no part in the conductivity. These are excess or N type conductors. The conductivity depends on the density of donotors, ΔE_1, and T.

Similarly some other impurity with negative valencies may have an energy state ΔE_2 volts above the top of the filled band. At room temperature or higher, an electron in the filled band may be raised in energy and

accepted by the impurity which then becomes a negative ion and usually is immobile. However, the resulting hole can take part in the conductivity.

In all cases represented in Fig. 6 an electron occupying a higher energy level than a positive ion or a hole has a certain probability that in any short interval of time it will drop into a lower energy state. However, during this same time interval there will be electrons which will be raised to a higher energy level by thermal agitation. When the number of electrons per second which are being elevated is equal to the number which are descending in energy, equilibrium prevails. The conductivity, σ, is then

$$\sigma = N e v_1 + P e v_2 \tag{9}$$

where N and P are the concentrations of electrons and holes respectively, e is the charge on the electron, v_1 and v_2 are the mobilities of electrons and holes respectively.

The above picture explains the following experimental facts which otherwise are difficult to interpret. 1) N type oxides, such as ZnO, when heated in a neutral or slightly reducing atmosphere become good conductors, presumably because they contain excess zinc which can donate electrons. If they then are heated in atmospheres which are increasingly more oxidizing their conductivity decreases until eventually they are intrinsic semiconductors or insulators. 2) P type oxides, such as NiO, when heat treated in strongly oxidizing atmospheres are good conductors. Very likely they contain oxygen in excess of the stoichiometric relation and this oxygen accepts additional electrons. When these are heated in less oxidizing or neutral atmospheres they become poorer conductors, semiconductors, or insulators. 3) When a P type oxide is sintered with another P type oxide, the conductivity increases. Similarly for two N type oxides. But when a P type is added to an N type the conductivity decreases. 4) If a metal forms several oxides the one in which the metal exerts its highest valence is N type, while the one in which it exerts its lowest valence will be P type.[5]

For several reasons it is desirable to survey the whole field of semiconductors for resistivity and temperature coefficient. One way in which this might be done is to draw a line in Figure 3 for each specimen. Before long such a figure would consist of such a maze of intersecting lines that it would be difficult to single out and follow any one line. The information can be condensed by plotting log ρ_0 versus B in equation (1) for each specimen.[6] The most important characteristics of a specimen thus are represented by a single point and many more specimens can be surveyed in a single diagram. Figure 7 shows such a plot for a large number of semiconductors investigated at these Laboratories or reported in the literature. Values for ρ_0 and B are given for $T = 25$ degrees centigrade. The points form a sort of

milky way. Semiconductors having a high ρ_0 are likely to have a high value of B and vice versa. If a series of semiconductors have points in Fig. 7 which fall on a straight line with a slope of $1/2.3T_0$, they have a common intercept in Fig. 3 for $(1/T) = 0$.

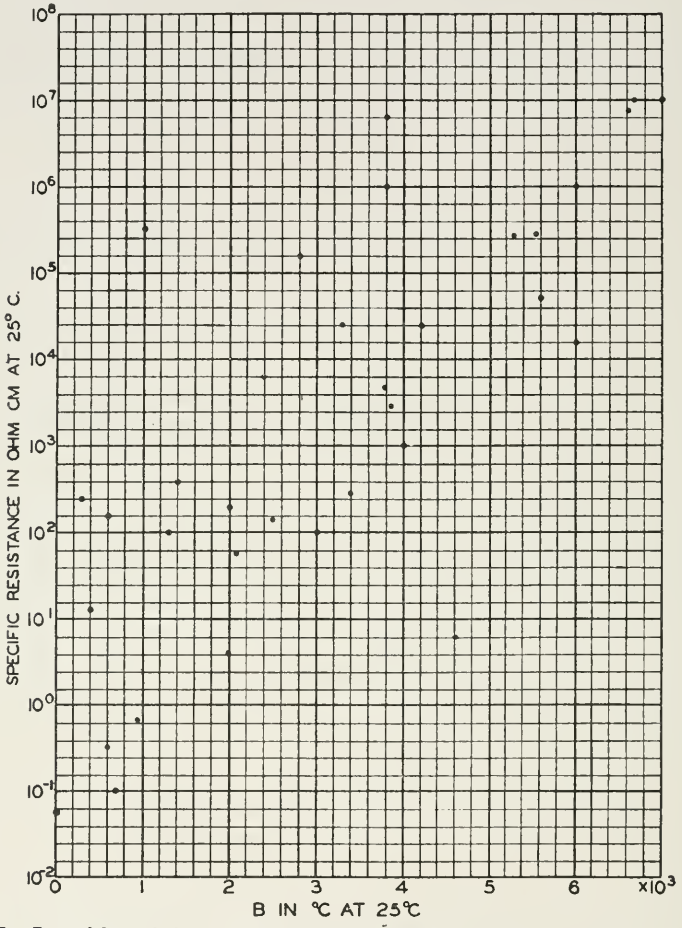


Fig. 7.—Logarithm of the resistivity of various semiconducting materials as a function of B in equation (1). The quantity, B, is proportional to the temperature coefficient of resistance as given in equation (4).

Physical Properties of Thermistors

One of the most interesting and useful properties of a thermistor is the way in which the voltage, V, across it changes as the current, I, through it increases. Figure 8 shows this relationship for a 0.061 centimeter diameter bead of Material No. 1 suspended in air. Each time the current is

changed, sufficient time is allowed for the voltage to attain a new steady value. Hence this curve is called the steady state curve. For sufficiently small currents, the power dissipated is too small to heat the thermistor appreciably, and Ohm's law is followed. However, as the current assumes larger values, the power dissipated increases, the temperature rises above ambient temperature, the resistance decreases, and hence the voltage is less than it would have been had the resistance remained constant. At some current, I_m, the voltage attains a maximum or peak value, V_m. Beyond

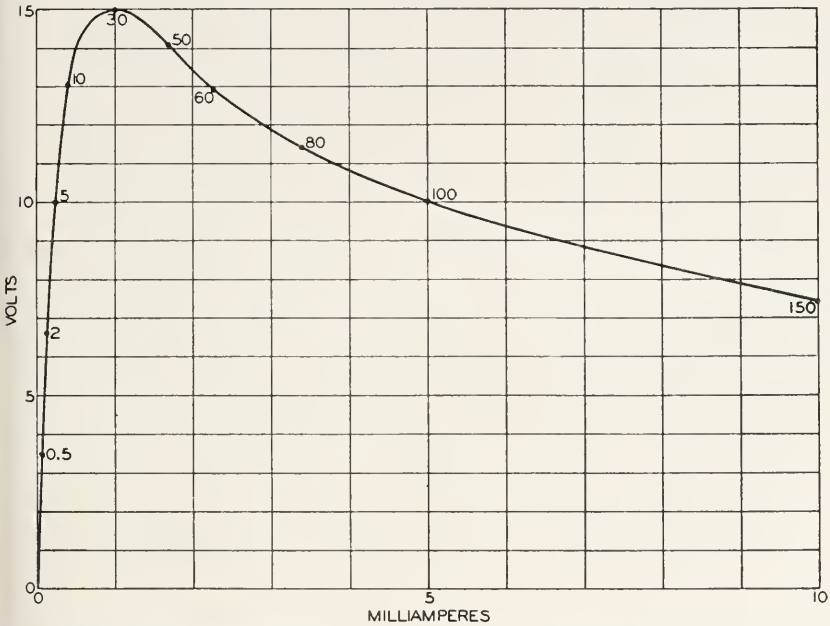


Fig. 8.—Static voltage-current curve for a typical thermistor. The numbers on the curve are the centigrade degrees rise in temperature above ambient.

this point as the current increases the voltage decreases and the thermistor is said to have a negative resistance whose value is dV/dI. The numbers on the curve give the rise in temperature above ambient temperature in centigrade degrees.

Because currents and voltages for different thermistors cover such a large range of values it has been found convenient to plot log V versus log I. Figure 9 shows such a plot for the same data as in Fig. 8. For various points on the curve, the temperature rise above ambient temperature is given. In a log plot, a line with a slope of 45 degrees represents a constant resistance; a line with a slope of -45 degrees represents constant power.

For a particular thermistor, the position of the log V versus log I plot is shifted, as shown in Fig. 10, by changing the dissipation constant C. This

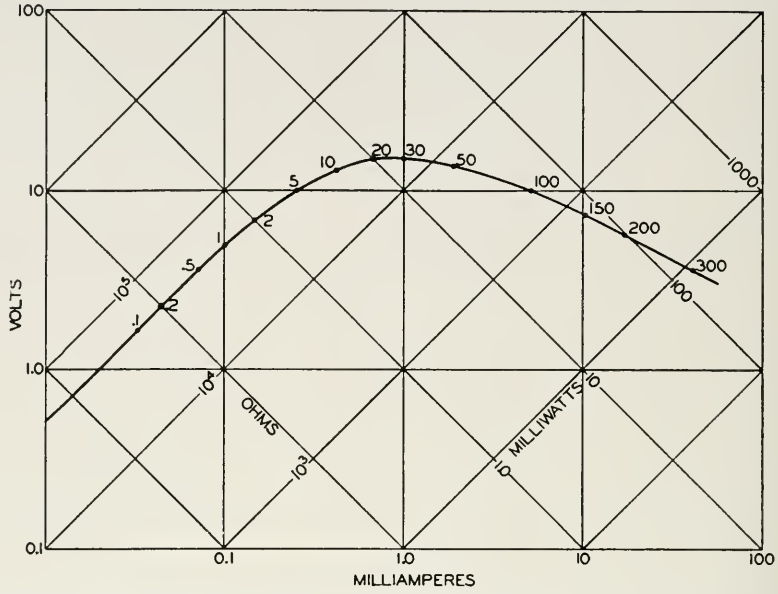


Fig. 9.—Logarithmic plot of static voltage-current curve for the same data as in Figure 8. The diagonal lines give the values of resistance and power.

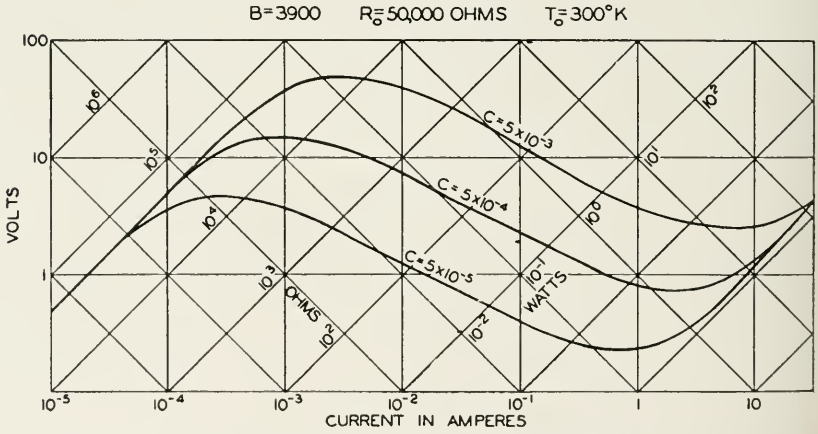


Fig. 10.—Logarithmic plots of voltage versus current for three values of the dissipation constant C. These curves are calculated for the constants given in the upper part of the figure.

can be done by changing the air pressure surrounding the bead, changing the medium, or changing the degree of thermal coupling between the thermis-

tor and its surroundings. The value of C for a particular thermistor in given surroundings can readily be determined from the V versus I curve in either Figs. 8 or 9. For each point, V/I is the resistance while V times I is W, the watts dissipated. The resistance data are converted to temperature from R versus T given by equation (2). A plot is then made of W versus T. For thermistors in which most of the heat is conducted away, W will increase linearly with T, so that C is constant. For thermistors suspended by fine wires in a vacuum, W will increase more rapidly than proportional to T, and C will increase with T. For thermistors of ordinary size and shape, in still air, C/Area = 1 to 40 milliwatts per centigrade degree per square centimeter depending upon the size and shape factor.

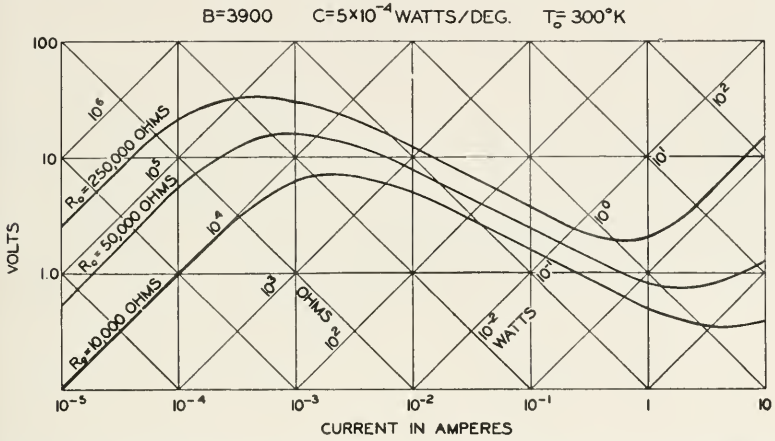


Fig. 11.—Logarithmic plots of voltage versus current for three values of the resistance, R_0, at ambient temperature. These curves are calculated for the constants given in the upper part of the figure.

The user of a thermistor may want to know how many watts can be dissipated before the resistance decreases by one per cent. This may be called the power sensitivity. It is equal to $C/(\alpha \times 100)$, and amounts to about one to ten milliwatts per square centimeter of area in still air. Both C and the power sensitivity increase with air velocity. The dependence of C on gas pressure and velocity is the basis of the use of thermistors as manometers and as anemometers or flowmeters. Note that in Fig. 10 one curve can be superposed on any other by a shift along a constant resistance line.

Figure 11 shows a family of log V versus log I curves for various values on R_0 while B, C, and T_0 are kept constant. This can be brought about by changing the length, width and thickness to vary R_0 while the surface area is kept constant. If the resistance had been changed by changing the ambient temperature, T_0, the resulting curves would not appear very different

from those shown. Note that one curve can be superposed on any other curve by a shift along a constant power line.

Figure 12 shows a family of log V versus log I curves for eight different values of B while C, R_0, and T_0 are kept constant. In contrast to the curves in Figs. 10 and 11 in which any curve could be obtained from any other curve by a shift along an appropriate axis, the curves in Fig. 12 are each distinct. For each curve there exists a limiting ohmic resistance for low

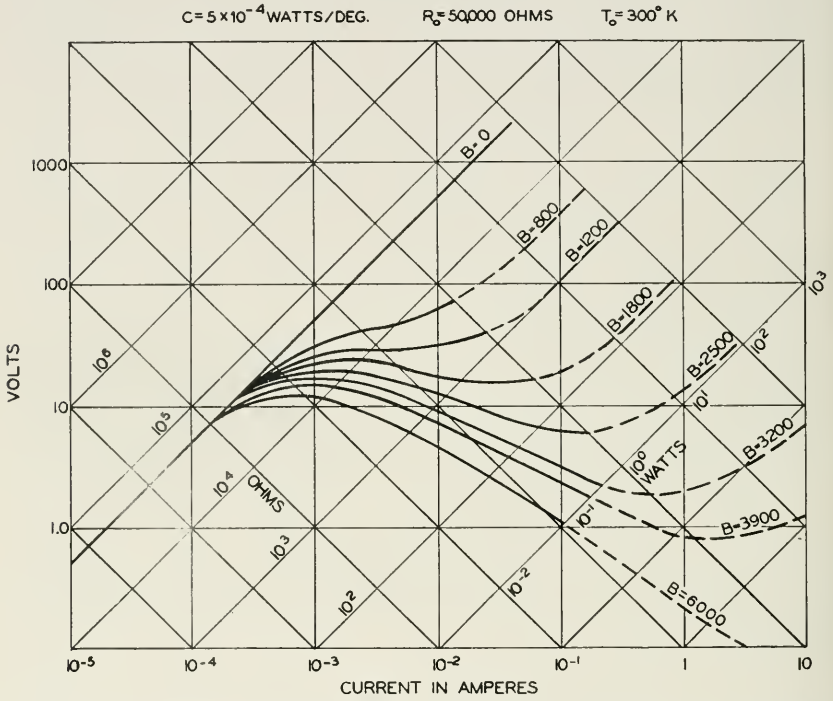


Fig. 12.—Logarithmic plots of voltage versus current for eight values of B in equation (1). These curves are calculated for the constants given in the upper part of the figure.

currents and another for high currents. For $B = 0$ these two are identical. As B becomes larger the log of the ratio of the two limiting resistances increases proportional to B. Note also that for $B > 1200\ K°$, the curves have a maximum. For large B values this maximum occurs at low powers and hence at low values of $T - T_0$. This follows since $W = C(T - T_0)$. As B decreases, V_m occurs at increasingly higher powers or temperatures. For $B < 1200\ K°$, no maximum exists.

The curves in Figs. 10 to 12 have been drawn for the ideal case in which the resistance in series with the thermistor is zero and in which no temperature limitations have been considered. In any actual case there is always

some unavoidable small resistance, such as that of the leads, in series with the thermistor and hence the parts of the curves corresponding to low resistances may not be observable. Also at high powers the temperature may attain such values that something in the thermistor structure will go to pieces thus limiting the range of observation. These unobservable ranges have been indicated by dashed lines in Fig. 12. The exact location of the dashed portions will of course depend on how a completed thermistor is constructed. In setting these limits consideration is given to temperature limitations beyond which aging effects might become too great.

The curves in Figs. 9 to 12 have been computed on the basis of the following equations:

$$R = R_0 \, \epsilon^{(B/T)-(B/T_0)} = V/I \tag{10}$$

$$W = C(T - T_0) = VI \tag{11}$$

For these curves the constants R_0, T_0, B, and C are specified. The values of temperature, T_m, power, W_m, resistance, R_m, voltage, V_m, and current, I_m, that prevail at the maximum in the voltage current curve are given by the following equations in which T_m is chosen as the independent parameter. By differentiating equations (10) and (11) with respect to I, putting the derivatives equal to zero, one obtains

$$T_m^2 = B(T_m - T_0) \tag{12}$$

whose solution is

$$T_m = (B/2)\,(1 \mp \sqrt{1 - 4T_0/B}). \tag{13}$$

The minus sign pertains to the maximum in Figs. 10 to 12 while the plus sign pertains to the minimum. Note that T_m depends only on B and T_0, and not on R, R_0 or C. From equations (4), (10) and (11) it follows that:

$$-\alpha_m\,(T_m - T_0) = 1 \tag{14}$$

$$W_m = C(T_m - T_0) \tag{15}$$

$$R_m = R_0 \, \epsilon^{-T_m/T_0} \doteq R_0 \, \epsilon^{-1}[1 - (T_m - T_0)/T_0 + (1/2)\{(T_m - T_0)/T_0\}^2 - \cdots] \tag{16}$$

$$V_m = [C\,R_0\,(T_m - T_0)\,(\epsilon^{-T_m/T_0})]^{1/2} = \{\{C\,R_0\,(T_m - T_0)\,\epsilon^{-1}\,[1 - (T_m - T_0)/T_0 + (1/2)\,\{(T_m - T_0)/T_0\}^2 - \cdots]\}\}^{1/2} \tag{17}$$

$$I_m = [(C/R_0)\,(T_m - T_0)\,\epsilon^{T_m/T_0}]^{1/2} = \{\{(C/R_0)\,(T_m - T_0)\,\epsilon[1 + (T_m - T_0)/T_0 + (1/2)\{(T_m - T_0)/T_0\}^2 + \cdots]\}\}^{1/2} \tag{18}$$

Thus far the presentation has been limited to steady state conditions, in which the power supplied to the thermistor is equal to the power dissipated by it, and the temperature remains constant. In many cases, however, it is important to consider transient conditions when the temperature, and any quantities which are functions of temperature, vary with time. A simple case which will illustrate the concepts and constants involved in such problems is as follows: A massive thermistor is heated to about 150 to 200 degrees centigrade by operating it well beyond the peak of its voltage

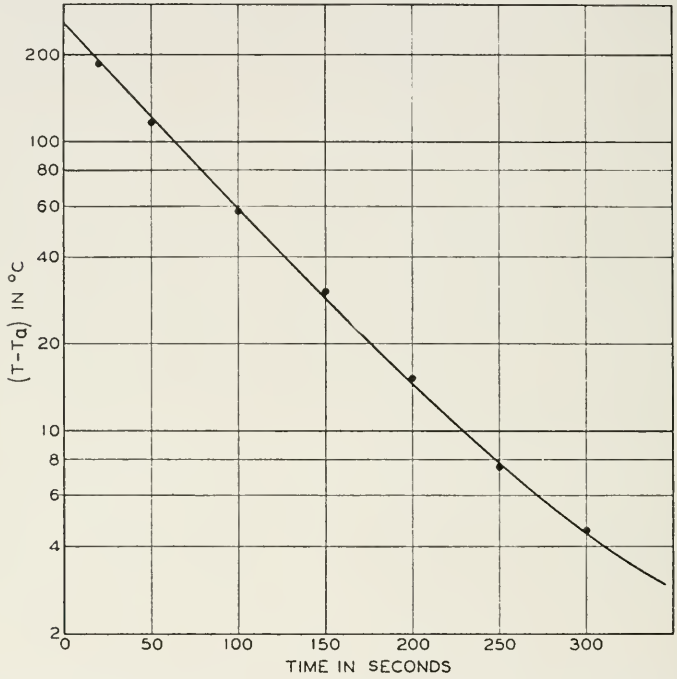


Fig. 13.—Cooling characteristic of a massive thermistor: log of temperature above ambient versus time.

current characteristic. At time $t = 0$, the circuit is switched over to a constant current having a value so small that I^2R is always negligibly small. The voltage across the thermistor is then followed as a function of time. From this, the resistance and temperature are computed. Figure 13 shows a plot of log $(T - T_a)$ versus t for a rod thermistor of Material No. 1 about 1.2 centimeters long, 0.30 centimeter in diameter and weighing 0.380 gram. In any time interval Δt, there are $C(T - T_a)\,\Delta t$ joules being dissipated. As a result the temperature will decrease by ΔT given by

$$-H\Delta T = C(T - T_a)\,\Delta t \text{ or } (T - T_a) = -(H/C)\,(\Delta T/\Delta t) \quad (19)$$

where H = heat capacity in joules per centigrade degree. The solution of this equation is

$$(T - T_a) = (T_0 - T_a)\,\epsilon^{-t/\tau} \tag{20}$$

in which $T_0 = T$ when $t = 0$ and

$$\tau = H/C, \tag{21}$$

where τ is in seconds. It is commonly called the time constant. From equation (20) it follows that a plot of log $(T - T_a)$ versus t should yield a straight line whose slope $= -1/2.303\tau$. If H and C vary slightly with temperature then τ will vary slightly with T and t. The line will not be perfectly straight but its slope at any t or $(T - T_a)$ will yield the appro-

TABLE I.—VALUES OF C, τ, H AS FUNCTIONS OF T FOR A THERMISTOR OF MATERIAL NO. 1 ABOUT 1.2 CENTIMETERS LONG, 0.30 CENTIMETERS IN DIAMETER AND WEIGHING 0.380 GRAM

T_a = 24 degrees centigrade

T Degrees Centigrade	C Watts per C. degree	τ Seconds	H Joules per C. degree	h Joules per gram per C. degree
44	0.0037	76	0.28	0.75
64	0.0037	74	0.27	0.72
84	0.0038	71	0.27	0.71
104	0.0037	69	0.26	0.68
124	0.0038	68	0.26	0.67
144	0.0038	67	0.26	0.67
164	0.0039	67	0.26	0.69
184	0.0041	66	0.27	0.71
204	0.0042	66	0.28	0.73

priate τ or H/C for that T. As previously described, C can be determined from a plot of watts dissipated versus T. For this thermistor this curve became steeper at the higher temperatures so that C increased for higher temperatures. Table I summarizes the values of C, τ, and H at various T for the unit in air.

When a thermistor is heated by passing current through it, conditions are somewhat more involved since the I^2R power will be a function of time. At any time in the heating cycle the heat power liberated will be equal to the watts dissipated or $C(T - T_a)$ plus watts required to raise the temperature or HdT/dt. The heat power liberated will depend on the circuit conditions. In a circuit like that shown in the upper corner of Figure 14, the current varies with time as shown by the six curves for six values of the battery voltage E. If a relay in the circuit operates when the current reaches a definite value, a considerable range of time delays can be achieved.

This family of curves will be modified by changes in ambient temperature and where rather precise time delays are required, the ambient temperature must be controlled or compensated.

Since thermistors cover a wide range in size, shape, and heat conductivity of surrounding media, large variations in H, C, and τ can be produced. The time constant can be varied from about one millisecond to about ten minutes or a millionfold.

One very important property of a thermistor is its aging characteristic or how constant the resistance at a given temperature stays with use. To obtain a stable thermistor it is necessary to: 1) select only semiconductors which are pure electronic conductors; 2) select those which do not change chemically when exposed to the atmosphere at elevated temperatures;

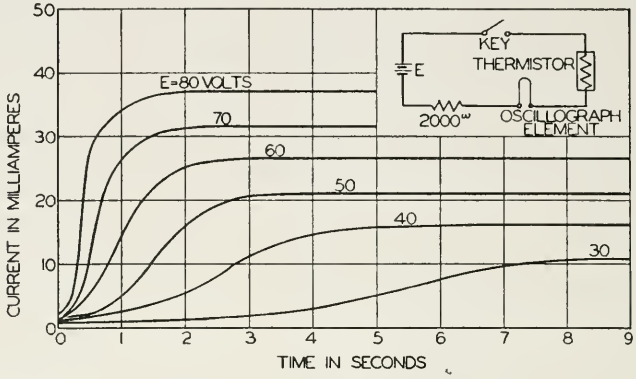


Fig. 14.—Current versus time curves for six values of the battery voltage in the circuit shown in the insert.

3) select one which is not sensitive to impurities likely to be encountered in manufacture or in use; 4) treat it so that the degree of dispersion of the critical impurities is in equilibrium or else that the approach to equilibrium is very slow at operating temperatures; 5) make a contact which is intimate, sticks tenaciously, has an expansion coefficient compatible with the semiconductor, and is durable in the atmospheres to which it will be exposed; 6) in some cases, enclose the thermistor in a thin coat of glass or material impervious to gases and liquids, the coat having a suitable expansion coefficient; 7) preage the unit for several days or weeks at a temperature somewhat higher than that to which it will be subjected. By taking these precautions remarkably good stabilities can be attained.

Figure 15 shows aging data taken on three-quarter inch diameter discs of Materials No. 1 and No. 2 with silver contacts and soldered leads. These discs were measured soon after production, were aged in an oven at 105 degrees centigrade and were periodically tested at 24 degrees centigrade.

The percentage change in resistance over its initial value is plotted versus the logarithm of the time in the aging oven. It is to be noted that most of the aging takes place in the first day or week. If these discs were preaged for a week or a month and the subsequent change in resistance referred to the resistance after preaging, they would age only about 0.2 per cent in one year. In a thermistor thermometer, this change in resistance would correspond to a temperature change of 0.05 centigrade degree. Thermistors mounted in an evacuated tube or coated with a thin layer of glass age even less than those shown in the figure. For some applications such high stability is not essential and it is not necessary to give the thermistors special treatment.

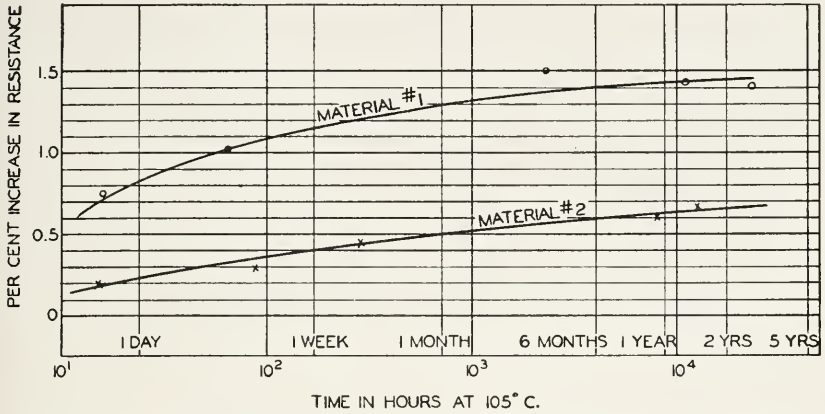


Fig. 15.—Aging characteristics of thermistors made of Materials No. 1 and No. 2 aged in an oven at 105°C. Per cent increase in resistance over its initial value versus time on a logarithmic scale.

Thermistors have been used at higher temperatures with satisfactory aging characteristics. Extruded rods of Material No. 1 have been tested for stability by treating them for two months at a temperature of 300 degrees centigrade. Typical units aged from 0.5 to 1.5 per cent of their initial resistance. Similar thermistors have been exposed alternately to temperatures of 300 degrees centigrade and −75 degrees centigrade for a total of 700 temperature cycles, each lasting one-half hour. The resistance of typical units changed by less than one per cent.

In some applications of thermistors very small changes in temperature produce small changes in potential across the thermistor which then are amplified in high gain amplifiers. If at the same time the resistance is fluctuating randomly by as little as one part in a million, the potential across the thermistor will also fluctuate by a magnitude which will be

directly proportional to the current. This fluctuating potential is called noise and since it depends on the current it is called current noise. In order to obtain the best signal to noise ratio, it is necessary that the current noise at operating conditions be less than Johnson or thermal noise.[7,8] To make noise-free units it is necessary to pay particular attention to the raw materials, the degree of sintering, the grain size, the method of making contact and any steps in the process which might result in minute surface cracks or fissures.

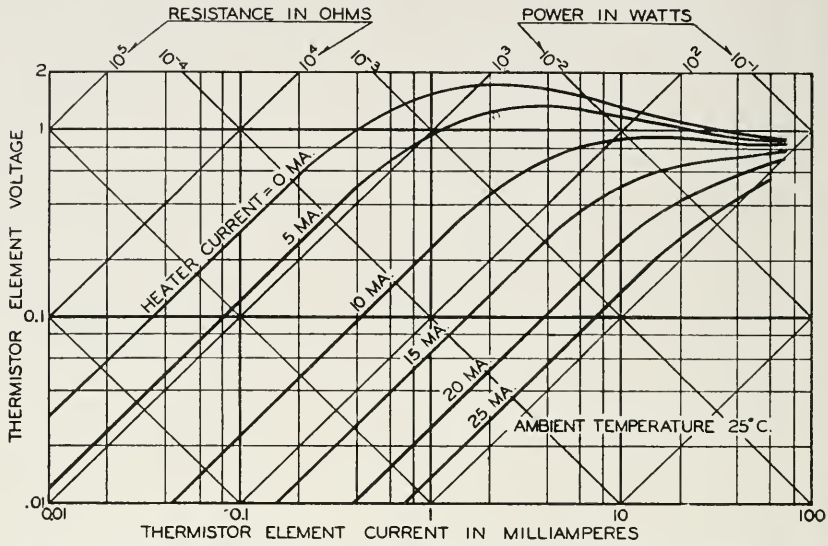


Fig. 16.—Logarithmic plots of voltage versus current for six values of heater curren in an indirectly heated thermistor. Resistance and power scales are given on the diag onal lines.

All the thermistors discussed thus far were either directly heated by the current passing through them or by changes in ambient temperature. In indirectly heated thermistors, the temperature and resistance of the thermistor are controlled primarily by the power fed into a heater thermally coupled to it. One such form might consist of a 0.038 centimeter diameter bead of Material No. 2 embedded in a small cylinder of glass about 0.38 centimeter long and 0.076 centimeter in diameter. A small nichrome heater coil having a resistance of 100 ohms is wound on the glass and is fused onto it with more glass. Figure 16 shows a plot of log V versus log I for the bead element at various currents through the heater. In this way the bead resistance can be changed from 3000 ohms to about 10 ohms. Indirectly heated thermistors are ordinarily used where the controlled circuit must be isolated electrically from the actuating circuit, and where the power from the latter must be fed into a constant resistance heater.

PART II—USES OF THERMISTORS

The thermistor, or thermally sensitive resistor, has probably excited more interest as a major electric circuit element than any other except the vacuum tube in the last decade. Its extreme versatility, small size and ruggedness were responsible for its introduction in great numbers into communications circuits within five years after its first application in this field. The next five year period spanned the war, and saw thermistors widely used in additional important applications. The more important of these uses ranged from time delays and temperature controls to feed-back amplifier automatic gain controls, speech volume limiters and superhigh frequency power meters. It is surprising that such versatility can result from a temperature dependent resistance characteristic alone. However, this effect produces a very useful nonlinear volt-ampere relationship. This, together with the ability to produce the sensitive element in a wide variety of shapes and sizes results in applications in diverse fields. The variables of design are many and interrelated, including electrical, thermal and mechanical dimensions.

The more important uses of thermistors as indication, control and circuit elements will be discussed, grouping the uses as they fall under the primary characteristics: resistance-temperature, volt-ampere, and current-time or dynamic relations.

Resistance-Temperature Relations

It has been pointed out in Part I that the temperature coefficient of electrical resistance of thermistors is negative and several times that of the ordinary metals at room temperature. In Thermistor Material No. 1, which is commonly used, the coefficient at 25 degrees centigrade is -4.4 per cent per centigrade degree, or over ten times that of copper, which is $+0.39$ per cent per centigrade degree at the same temperature. A circuit element made of this thermistor material has a resistance at zero degrees centigrade which is nine times the resistance of the same element at 50 degrees centigrade. For comparison, the resistance of a copper wire at 50 degrees centigrade is 1.21 times its value at zero degrees centigrade.

The resistance-temperature characteristics of thermistors suggest their use as sensitive thermometers, as temperature actuated controls and as compensators for the effects of varying ambient temperature on other elements in electric circuits.

Thermometry

The application of thermistors to temperature measurement follows the usual principles of resistance thermometry. However, the large value of temperature coefficient of thermistors permits a new order of sensitivity to be obtained. This and the small size, simplicity and ruggedness of thermis-

tors adapt them to a wide variety of temperature measuring applications. When designed for this service, thermistor thermometers have long-time stability which is good for temperatures up to 300 degrees centigrade and excellent for more moderate temperatures. A well aged thermistor used in precision temperature measurements was found to be within 0.01 centigrade degree of its calibration after two months use at various temperatures up to 100 degrees centigrade. As development proceeds, the stability of thermistor thermometers may be expected to approach that of precision platinum thermometers. Conventional bridge or other resistance measuring circuits are commonly employed with thermistors. As with any resistance thermometer, consideration must be given to keeping the measuring current sufficiently small so that it produces no appreciable heating in order that the

TABLE II.—TEMPERATURE-RESISTANCE CHARACTERISTIC OF A TYPICAL THERMISTOR THERMOMETER

Temperature	Resistance	Temperature Coefficients	
		B	α
−25°C.	580,000 ohms	3780 C. deg.	−6.1%/ C. deg.
0	145,000	3850	−5.2
25	46,000	3920	−4.4
50	16,400	3980	−3.8
75	6,700	4050	−3.3
100	3,200	4120	−3.0
150	830	4260	−2.4
200	305	4410	−2.0
275	100	4600	−1.5

Dissipation constant in still air, approx. ..... 4 milliwatts/C. deg.
Thermal time constant in still air, approx. ..... 70 seconds
Dimensions of thermistor, diameter approx. ..... 0.11 inch
length approx. ..... 0.54 inch

thermistor resistance shall be dependent upon the ambient temperature alone.

Since thermistors are readily designed for higher resistance values than metallic resistance thermometers or thermocouples, lead resistances are not ordinarily bothersome. Hence the temperature sensitive element can be located remotely from its associated measuring circuit. This permits great flexibility in application, such as for instance wire line transmission of temperature indications to control points.

Table II gives the characteristics of a typical thermistor thermometer. The dissipation constant is the ratio of the power input in watts dissipated in the thermistor to the resultant temperature rise in centigrade degrees. The time constant is the time required for the temperature of the thermistor to change 63 per cent of the difference between its initial value and that of the surroundings. As a sensitive thermometer, this thermistor with a simple Wheatstone bridge and a galvanometer whose sensitivity is 2 ×

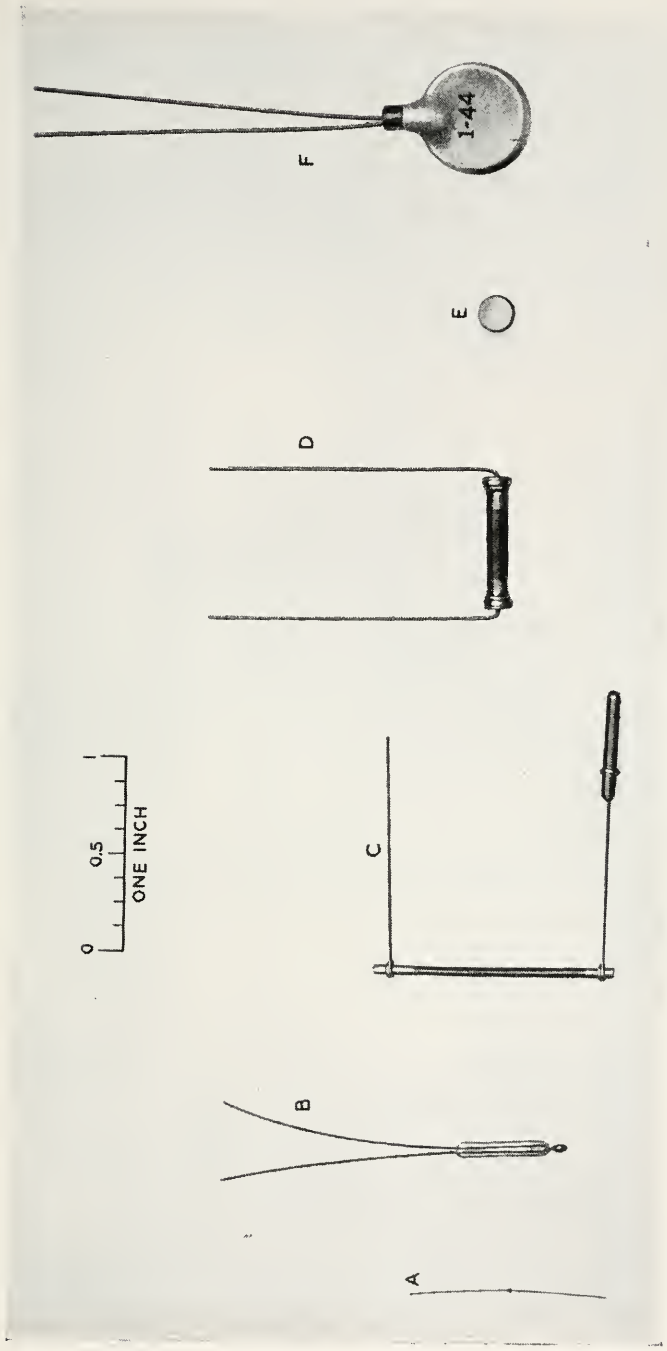


Fig. 17.—Some forms of thermistors which have been used as resistance thermometers.

10^{-10} amperes per millimeter per meter will readily indicate a temperature change of 0.0005 centigrade degree. For comparison a precision platinum resistance thermometer and the required special bridge such as the Mueller will indicate a minimum change of 0.003 centigrade degree with a similar galvanometer.

Several thermistors which have been used for thermometry are shown in Fig. 17. Included in the group are types which are suited to such diverse applications as intravenous blood thermometry and supercharger rotor temperature measurement. In Fig. 17, A is a tiny bead with a response time of less than a second in air. B is a probe type unit for use in air streams or liquids. C is a meteorological thermometer used in automatic radio transmission of weather data from free balloons. D is a rod shaped unit. E is a disc or pellet, adapted for use in a metal thermometer bulb. Discs like the one shown have been sweated to metal plates to give a low thermal impedance connection to the object whose temperature is to be determined. F is a large disc with an enveloping paint finish for use in humid surroundings. The characteristics of these types are given in Table III.

The temperature of objects which are inaccessible, in motion, or too hot for contact thermometry can be determined by permitting radiation from the object to be focussed on a suitable thermistor by means of an elliptical mirror. Such a thermistor may take the form of a thin flake attached to a solid support. Its advantages compared with the thermopile and resistance bolometer are its more favorable resistance value, its ruggedness, and its high temperature coefficient of resistance. It can be made small to reduce its heat capacity so as rapidly to follow changing temperatures. Flake thermistors have been made with time constants from one millisecond to one second. Since the amount of radiant power falling on the thermistor may be quite small, sensitive meters or vacuum tube amplifiers are required to measure the small changes in the flake resistance. Where rapidly varying temperatures are not involved, thermistors with longer time constants and simpler circuit equipments can be utilized.

Temperature Control

The use of thermistors for temperature control purposes is related closely to their application as temperature measuring devices. In the ideal temperature sensitive control element, sensitivity to temperature change should be high and the resistance value at the control temperature should be the proper value for the control circuit used. Also the temperature rise of the control element due to circuit heating should be low, and the stability of calibration should be good. The size and shape of the sensitive element are dictated by several factors such as the space available, the required speed of response to temperature changes and the amount of power which must

be dissipated in the element by the control circuit to permit the arrangement to operate relays, motors or valves.

Because of their high temperature sensitivity, thermistors have shown much promise as control elements. Their adaptability and their stability at relatively high temperatures led, for instance, to an aircraft engine control system using a rod-shaped thermistor as the control element.[9] The

TABLE III.—THERMISTOR THERMOMETERS

	A	B	C	D	E	F
Nominal Resistance, Ohms at						
−25°C	—	—	87,500	610,000	—	13,000
0	5,000	325,000	37,500	153,000	490	3,200
25	2,000	100,000	18,000	48,500	175	950
50	900	33,000	9,700	17,300	71	340
75	460	13,000	5,500	7,100	32	145
100	250	6,000	3,700	3,400	16	70
150	95	1,600	—	870	4.5	—
200	—	500	—	—	1.6	—
300	—	80	—	—	—	—
Temp. Coeff. α, %/C. deg. at 25°C	−3.4	−4.4	−2.8	−4.4	−3.8	−4.4
Max. Permissible Temp., °C.	150	300	100	150	200	100
Dissipation Constant, C, mw/C. deg.						
Still air	0.1	1	7	7	—	20
Still water	—	7	—	—	—	—
Thermal Time Constant, Seconds						
Still air	1	30	25	60	—	—
Still water	—	4	—	—	—	—
Shape	Bead	Probe	Rod	Rod	Disc	Disc
Dimensions, Inches						
Diameter or Width	0.015	0.1	0.05	0.15	0.2	0.56
Length or Thickness (less leads)	0.02	0.6	1.2	0.7	0.1	0.31

thermistor, mounted in a standard one-quarter inch diameter temperature bulb assembly, operated at approximately 275 degrees centigrade. It was associated with a differential relay and control motor on the aircraft 28 volt d-c system. The power dissipation in the thermistor was two watts. The resistance of a typical thermistor under these high temperature conditions remained within ± 1.5 per cent over a period of months. This corresponds to about ± one centigrade degree variation in calibration. Several other related designs were developed using the same control system

with other thermistors designed for both higher and lower temperature operation. In the lower temperature applications, typical thermistors maintained their calibrations within a few tenths of a centigrade degree.

In general, electron tube control circuits dissipate less power in the thermistor than relay circuits do. This results in less temperature rise in the thermistor and leads to a more accurate control. While the average value of this temperature rise can be allowed for in the design, the variations in different installations require individual calibration to correct the errors if they are large. The corrections may be different as a result of variations of the thermal conductivity of the surrounding media from time to time or from one installation to another. The greater the power dissipated in the thermistor the greater the absolute error in the control temperature for a given change in thermal conductivity. This follows from the relation

$$\Delta T = W/C \tag{22}$$

where ΔT is the temperature rise, W is the power dissipated and C is the dissipation constant which depends on thermal coupling to the surroundings. For the same reason, the temperature indicated by a resistance thermometer immersed in an agitated medium will depend on the rate of flow if the temperature sensitive element is operated several degrees hotter than its surroundings.

The design of a thermistor for a ventilating duct thermostat might proceed as follows as far as temperature rise is concerned:

1. Determine the power dissipation. This depends upon the circuit selected and the required overall sensitivity.

2. Estimate the permissible temperature rise of the thermistor, set by the expected variation in air speed and the required temperature control accuracy.

3. Solve Equation (22) for the dissipation constant and select a thermistor of appropriate design and size for this constant in the nominal air speed. Where more than one style of thermistor is available, the required time constant will determine the choice.

Compensators

It is a natural and obvious application of thermistors to use them to compensate for changes in resistance of electrical circuits caused by ambient temperature variations. A simple example is the compensation of a copper wire line, the resistance of which increases approximately 0.4 per cent per centigrade degree. A thermistor having approximately one-tenth the resistance of the copper, with a temperature coefficient of -4 per cent per centigrade degree placed in series with the line and subjected to the same ambient temperature, would serve to compensate it over a narrow tempera-

ture range. In practice however, the compensating thermistor is associated with parallel and sometimes series resistance, so that the combination gives a change in resistance closely equal and opposite to that of the circuit to be compensated over a wide range of temperatures. See Fig. 18.

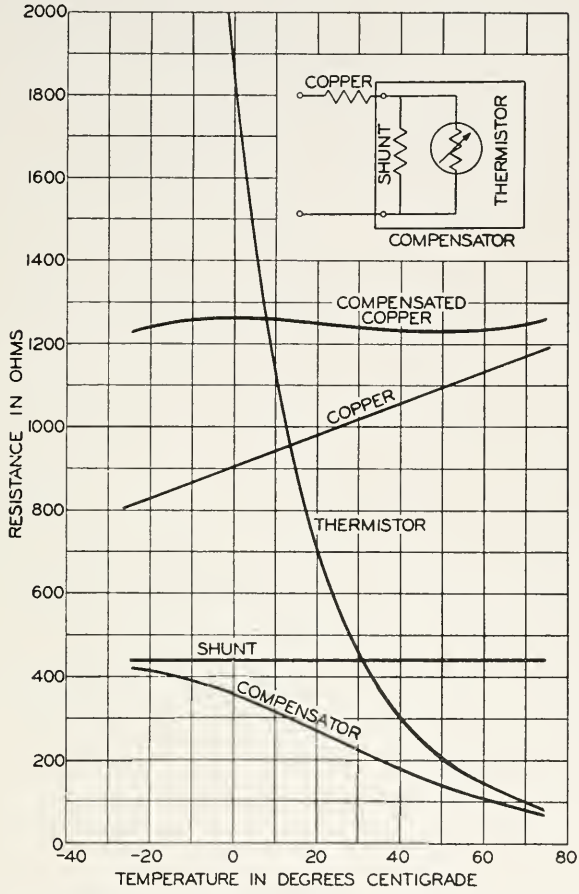


Fig. 18.—Temperature compensation of a copper conductor by means of a thermistor network.

A copper winding having a resistance of 1000 ohms at 25 degrees centigrade can be compensated by means of a thermistor of 566 ohms at 25 degrees centigrade in parallel with an ohmic resistance of 445 ohms as shown in Fig. 18. The winding with compensator has a resistance of 1250 ohms constant to ± 1.6 per cent over the temperature range −25 degrees centigrade to +75 degrees centigrade. Over this range the copper alone varies from 807.5 ohms to 1192.5 ohms, or ± 19 per cent about the mean. The

total resistance of the circuit has been increased only 6.1 per cent at the upper temperature limit by the addition of a compensator. This increase is small because of the high temperature coefficient of the compensating thermistor. The characteristics of such a thermistor are so stable that the resistance would remain constant within less than one per cent for ten years if maintained at any temperature up to about 100 degrees centigrade. Figure 15 shows aging characteristics for typical thermistors suitable for use in compensators. These curves include the change which occurs during the seasoning period of several days at the factory, so that the aging in use is a fraction of the total shown.

In many circuits which need to function to close tolerances under wide ambient temperature variation, the values of one or more circuit elements may vary undesirably with temperature. Frequently the resultant overall variation with temperature can be reduced by the insertion of a simple thermistor placed at an appropriate point in the circuit. This is particularly true if the circuit contains vacuum tube amplifiers. In this manner frequency and gain shifts in communications circuits have been cancelled and temperature errors prevented in the operation of devices such as electric meters. The change in inductance of a coil due to the variation of magnetic characteristics of the core material with temperature has been prevented by partially saturating the coil with direct current, the magnitude of which is directly controlled by the resistance of a thermistor imbedded in the core. In this way the amount of d-c magnetic flux is adjusted by the thermistor so that the inductance of the coil is independent of temperature.

In designing a compensator, care must be taken to ensure exposure of the thermistor to the temperature affecting the element to be compensated. Power dissipation in the thermistor must be considered and either limited to a value which will not produce a significant rise in temperature above ambient, or offset in the design.

Volt-Ampere Characteristics

The nonlinear shape of the static characteristic relating voltage, current, resistance and power for a typical thermistor was illustrated by Fig. 9. The part of the curve to the right of the voltage maximum has a negative slope, applicable in a large number of ways in electric circuits. The particular characteristic shown begins with a resistance of approximately 50,000 ohms at low power. Additional power dissipation raises the temperature of the thermistor element and decreases its resistance. At the voltage maximum the resistance is reduced to about one-third its cold value, or 17,000 ohms, and the dissipation is 13 milliwatts. The resistance becomes approximately 300 ohms when the dissipation is 100 milliwatts. Such resistance-power characteristics have resulted in the use of thermistors as sensitive power measuring devices, and as automatically variable resistances

for such applications as output amplitude controls for oscillators and amplifiers. Their nonlinear characteristics also fit thermistors for use as voltage regulators, volume controls, expandors, contactless switches and remote control devices. To permit their use in these applications for d-c as well as a-c circuits, nonpolarizing semiconductors alone are employed in thermistors with the exception of two early types.

Power Meter

Thermistors have been used very extensively in the ultra and superhigh frequency ranges in test sets as power measuring elements. The particular advantages of thermistors for this use are that they can be made small in size, have a small electrical capacity, can be severely overloaded without

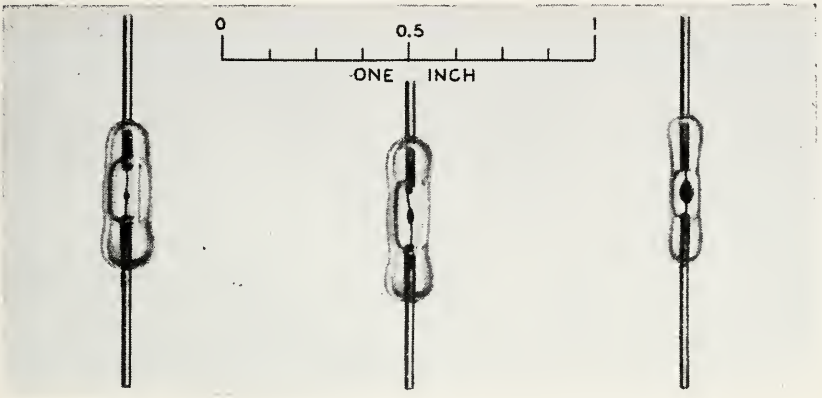


Fig. 19.—Power measuring thermistors with different sized beads.

change in calibration, and can easily be calibrated with direct-current or low-frequency power. For this application the thermistor is used as a power absorbing terminating resistance in the transmission line, which may be of Lecher, coaxial or wave-guide form. Methods of mounting have been worked out which reduce the reflection of high frequency energy from the termination to negligible values and assure accurate measurement of the power over broad bands in the frequency spectrum. Conventionally, the thermistor is operated as one arm of a Wheatstone bridge, and is biased with low frequency or d-c energy to a selected operating resistance value, for instance 125 or 250 ohms in the absence of the power to be measured. The application of the power to be measured further decreases the thermistor resistance, the bridge becomes unbalanced and a deflection is obtained on the bridge meter. A full scale power indication of one milliwatt is customary for the test set described, although values from 0.1 milliwatt to 200 milliwatts have been employed using thermistors with different sized beads as shown in Fig. 19.

Continuous operation tests of these thermistors indicate very satisfactory stability with an indefinitely long life. A group of eight power meter thermistors, normally operated at 10 milliwatts and having a maximum rating of 20 milliwatts, were operated for over 3000 hours at a power input of 30 milliwatts. During this time the room temperature resistance remained within 1.5 per cent of its initial value, and the power sensitivity, which is the significant characteristic, changed by less than 0.5 per cent.

When power measuring test sets are intended for use with wide ambient temperature variations, it is necessary to temperature compensate the thermistor. This is accomplished conventionally by the introduction of two other thermistors into the bridge circuit. These units are designed to be insensitive to bridge currents but responsive to ambient temperature. One of the compensators maintains the zero point and the other holds the meter scale calibration independent of the effect of temperature change on the measuring thermistor characteristics.

Automatic Oscillator Amplitude Control

Meacham,[10] and Shepherd and Wise[11] have described the use of thermistors to provide an effective method of amplitude stabilization of both low and high frequency oscillators. These circuits oscillate because of positive feedback around the vacuum tube. The feedback circuit is a bridge with at least one arm containing a thermistor which is heated by the oscillator output. Through this arrangement, the feedback depends in phase and magnitude upon the output, and there is one value of thermistor resistance which if attained would balance the bridge and cause the oscillation amplitude to vanish. Obviously this condition can never be exactly attained, and the operating point is just enough different to keep the bridge slightly unbalanced and produce a predetermined steady value of oscillation output. Such oscillators in which the amplitude is determined by thermistor nonlinearity have manifold advantages over those whose amplitude is limited by vacuum tube nonlinearity. The harmonic content in the output is smaller, and the performance is much less dependent upon the individual vacuum tube and upon variations of the supply voltages. It is necessary that the thermal inertia of the thermistor be sufficient to prevent it from varying in resistance at the oscillation frequency. This is easily satisfied for all frequencies down to a small fraction of a cycle per second. Figure 20 shows a thermistor frequently used for oscillator control together with its static electrical characteristic. This thermistor is satisfactory in oscillators for frequencies above approximately 100 cycles per second. Similar types have been developed with response characteristics suited to lower frequencies and for other resistance values and powers.

Where the ambient temperature sensitivity of the thermistor is disadvantageous in oscillator controls, the thermistor can be compensated by

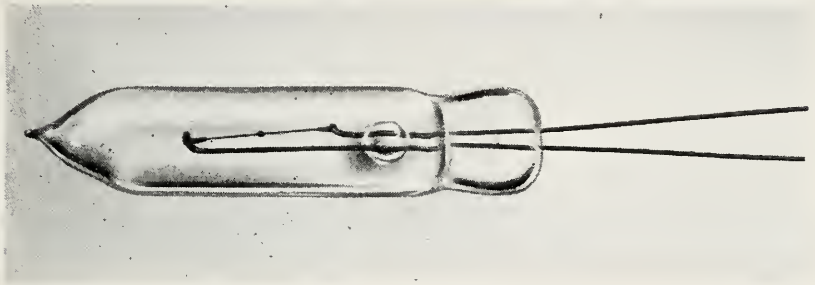


Fig. 20A.—An amplitude control thermistor. The glass bulb is 1.5 inches in length.

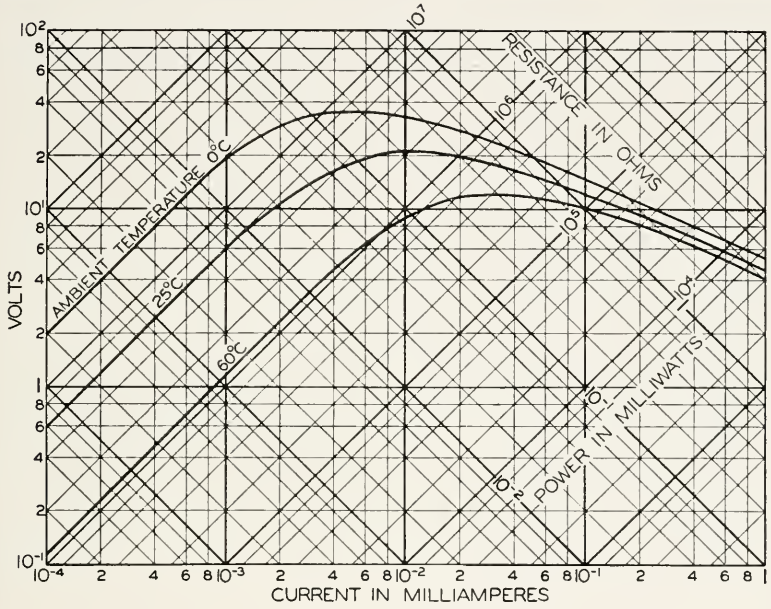


Fig. 20B.—Steady state characteristics of amplitude control thermistor shown in Figure 20A.

thermostating it with a heater and compensating thermistor network, as shown in Fig. 21.

Amplifier Automatic Gain Control

Since the resistance of a thermistor of suitable design varies markedly with the power dissipated in it or in a closely associated heater, such ther-

mistors have proven to be very valuable as automatic gain controls, especially for use with negative feedback amplifiers. This arrangement has seen extensive use in wire communication circuits for transmission level regulation, and has been described in some detail elsewhere.[12, 13, 14] In one form, a directly heated thermistor is connected into the feedback circuit of the amplifier in such a way that the amount of feedback voltage is varied to compensate for any change in the output signal. By this arrangement, the gain of each amplifier in the transmission system is continually adjusted to correct for variations in overall loss due to weather conditions and other factors, so that constant transmission is obtained over the channel at all times. In the Type K2 carrier system now in extensive use, the system gain is regulated principally in this way. In this system the transmission loss variations due to temperature are not the same in all parts of the pass band. The loss is corrected at certain repeater points along the transmission line by two additional thermistor gain controls: slope, proportional to fre-

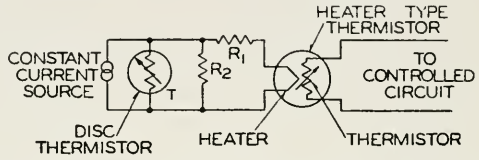


Fig. 21.—Circuit employing an auxiliary disc thermistor to compensate for effect of varying ambient temperature on a control thermistor.

quency, and bulge, with a maximum at one frequency. These thermistors are indirectly heated, with their heaters actuated by energy dependent upon the amplitude of the separate pilot carriers which are introduced at the sending end for the purpose.

In this type of application, the thermistor will react to the ambient temperature to which it is exposed, as well as to the current passing through it. Where this is important, the reaction to ambient temperature can be eliminated by the use of a heater type thermistor as shown in Fig. 21. The heater is connected to an auxiliary circuit containing a temperature compensating thermistor. This circuit is so arranged that the power fed into the heater of the gain control thermistor is just sufficient at any ambient temperature to give a controlled and constant value of temperature in the vicinity of the gain control thermistor element.

Another interesting form of thermistor gain control utilizes a heater type thermistor, with the heater driven by the output of the amplifier and with the thermistor element in the input circuit, as shown in Fig. 22. In this arrangement the feedback is accomplished by thermal, rather than electrical coupling. A broad-band carrier system, Type L1, is regulated

with this type of thermistor. In this system a pilot frequency is supplied, and current of this frequency, selected by a network in the regulator, actuates the heater of the thermistor to give smooth, continuous gain control.

By utilizing a heater thermistor of different characteristics, the circuit and load of Fig. 22 may be given protection against overloads. In this application the sensitivity and element resistance of the thermistor are chosen so that the thermistor element forms a shunt of high resistance value so as to have negligible effect on the amplifier for any normal value of output. However, if the output power rises to an abnormal level, the thermistor element becomes heated and reduced in resistance. This shunts the input to the amplifier and thus limits the output. Choice of a thermistor having a suitable time constant permits the onset of the limiting effect to be delayed for any period from about a second to a few minutes.

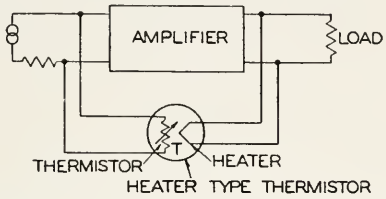


Fig. 22.—Thermal feedback circuit for gain control purposes. This arrangement has also been used as a protective circuit for overloads.

Regulators and Limiters

A group of related applications for thermistors depends on their steady state nonlinear volt-ampere characteristic. These are the voltage regulator, the speech volume limiter, the compressor and the expandor. The compressor and expandor are devices for altering the range of signal amplitudes. The compressor functions to reduce the range, while the expandor increases it. In Fig. 23, Curve 1 is a typical thermistor static characteristic having negative slope to the right of the voltage maximum. Curve 2 is the characteristic of an ohmic resistance R having an equal but positive slope. Curve 3 is the characteristic obtained if the thermistor and resistor are placed in series. It has an extensive segment where the voltage is almost independent of the current. This is the condition for a voltage regulator or limiter. If a larger value of resistance is used, as in Curve 4, its combination with the thermistor in series results in Curve 5, the compressor. In these uses the thermistor regulator is in shunt with the load resistance, so that in the circuit diagram of Fig. 23,

$$E = E_o = E_I - IR_S. \tag{23}$$

Here E is the voltage across the thermistor and resistor R, E_o is the output

voltage, and E_I, I and R_S are respectively the input voltage, current and resistance.

If the thermistor and associated resistor are placed in series between the generator and load resistance, an expandor is obtained, and

$$E_o = E_I - E. \tag{24}$$

As the resistance R in series with the thermistor is increased, the degree of expansion is decreased and vice versa.

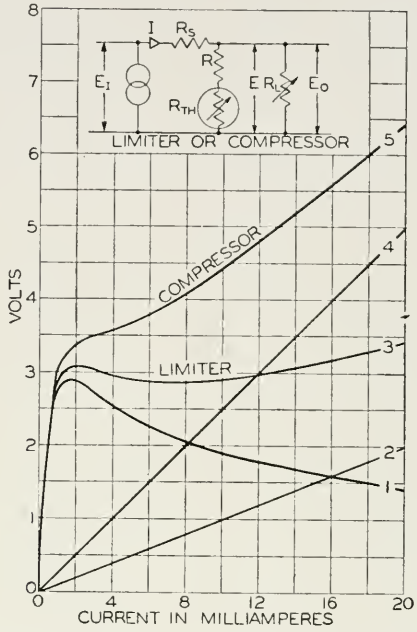


Fig. 23.—Characteristics of a simple thermistor voltage regulator, limiter or compressor circuit.

The treatment thus far in this section assumes that change of operating point occurs slowly enough to follow along the static curves. For a sufficiently rapid change of the operating point, the latter departs from the static curve and tends to progress along an ohmic resistance line intersecting the static curve. For sufficiently rapid fluctuations, control action may then be derived from the resistance changes resulting from the r.m.s. power dissipated in the thermistor unit. In speech volume limiters, the thermistor is designed for a speed of response that will produce limiting action for the changes in volume which are syllabic in frequency or slower, and that will not follow the more rapid speech fluctuations with resulting change in wave

shape or nonlinear distortion. Speech volume limiters of this type can accommodate large volume changes without producing wave form distortion.[13,15]

Remote Control Swiches

The contactless switch and rheostat are natural extensions of the uses just discussed. The thermistor is used as an element in the circuit which is to be controlled, while the thermistor resistance value is in turn dependent upon the energy dissipated directly or indirectly in it by the controlling circuit. By taking advantage of the nonlinearity of the static volt-ampere characteristic, it is possible to provide snap and lock-in action in some applications.

Manometer

Several interesting and useful applications such as vacuum gauges, gas analyzers, flowmeters, thermal conductivity meters and liquid level gauges of high sensitivity and low operating temperature are based upon the physical principle that the dissipation constant of the thermistor depends on the thermal conductivity of the medium in which it is immersed. As shown in Fig. 10, a change in this constant shifts the position of the static characteristic with respect to the axes. In these applications, the undesired response of the thermistor to the ambient temperature of the medium can in many cases be eliminated or reduced by introducing a second thermistor of similar characteristics into the measuring circuit. The compensating thermistor is subjected to the same ambient temperature, but is shielded from the effect being measured, such as gas pressure or flow. The two thermistors can be connected into adjacent arms of a Wheatstone bridge which is balanced when the test effect is zero and becomes unbalanced when the effective thermal conductivity of the medium is increased. In gas flow measurements, the minimum measurable velocity is limited, as in all "hot wire" devices, by the convection currents produced by the heated thermistor.

The vacuum gauge or manometer which is typical of these applications will be described somewhat in detail. The sensitive element of the thermistor manometer is a small glass coated bead 0.02 inch in diameter, suspended by two fine wire leads in a tubular bulb for attachment to the chamber whose gas pressure is to be measured. The volt-ampere characteristics of a typical laboratory model manometer are shown in Fig. 24 for air at several absolute pressures from 10^{-6} millimeters of mercury to atmospheric. The operating point is in general to the right of the peak of these curves. Electrically this element is connected into a unity ratio arm Wheatstone bridge with a similar but evacuated thermistor in an adjacent arm as shown in the circuit

schematic of Fig. 25. The air pressure calibration for such a manometer is also shown. The characteristic will be shifted when a gas is used having a thermal conductivity different from that of air. Such a manometer has been found to be best suited for the measurement of pressures from 10^{-5} to 10 millimeters of mercury. The lower pressure limit is set by practical considerations such as meter sensitivity and the ability to maintain the zero setting for reasonable periods of time in the presence of the variations of supply voltage and ambient temperature. The upper pressure measurement limit is caused by the onset of saturation in the bridge unbalance

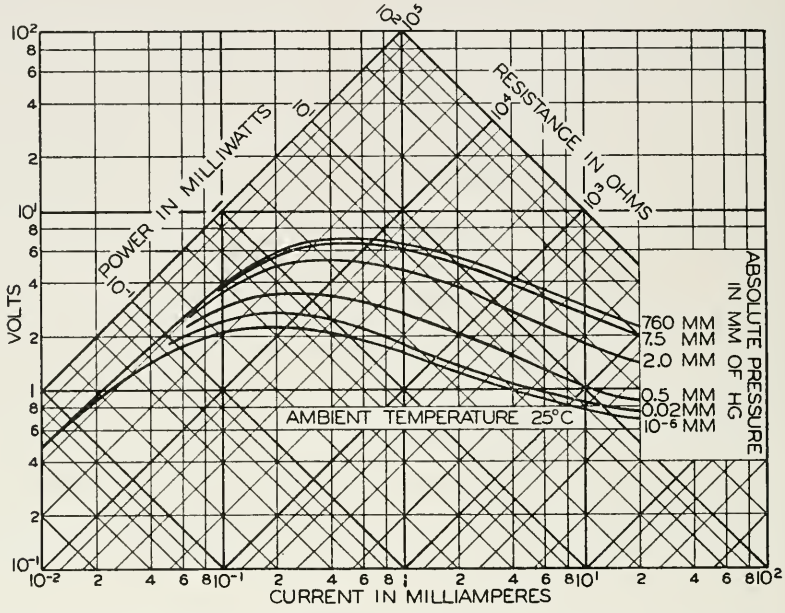


Fig. 24.—Characteristics of a typical thermistor manometer tube, showing the effect of gas pressure on the volt-ampere and resistance-power relations.

voltage versus pressure characteristic at high pressures. This is basically because the mean free path of the gas molecules becomes short compared with the distance between the thermistor bead and the inner surface of the manometer bulb, so that the cooling effect becomes nearly independent of the pressure.

The thermistor manometer is specially advantageous for use in gases which may be decomposed thermally. For this type of use, the thermistor element temperature can be limited to a rise of 30 centigrade degrees or less above ambient temperature. For ordinary applications, however, a temperature rise up to approximately 200 centigrade degrees in vacuum

permits measurement over wider ranges of pressure. Special models have also been made for use in corrosive gases. These expose only glass and platinum alloy to the gas under test.

Timing Devices

The numerically greatest application for thermistors in the communication field has been for time delay purposes. The physical basis for this use has

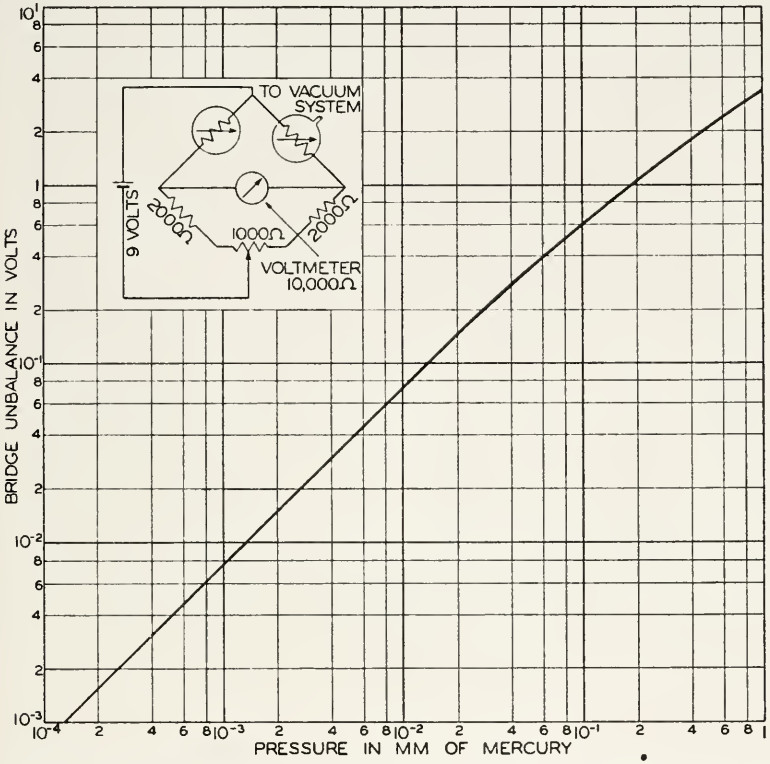


Fig. 25.—Operating circuit and calibration for a vacuum gauge utilizing the thermistor of Figure 24.

been discussed in Part I for the case of a directly heated thermistor placed in series with a voltage source and a load to delay the current rise after circuit closure. This type of operation will be termed the power driven time delay.

By the use of a thermistor suited to the circuit and operating conditions, power driven time delays can be produced from a few milliseconds to the order of a few minutes. Thermistors of this sort have the advantage of small size, light weight, ruggedness, indefinitely long life and absence of contacts, moving parts, or pneumatic orifices which require maintenance

care. Power driven time delay thermistors are best fitted for applications where close limits on the time interval are not required. In some communications uses it is satisfactory to permit a six to one ratio between maximum and minimum times as a result of the simultaneous variation from nominal values of all the following factors which affect the delay: operating voltage ± 5 per cent; ambient temperature 20 degrees centigrade to 40 degrees centigrade; operating current of the relay ± 25 per cent; relay resistance ± 5 per cent; and thermistor variations such as occur from unit to unit of the same type.

After a timing operation a power driven time delay thermistor should be allowed time to cool before a second operation. If this is not done, the second timing interval will be shorter than the first. The cooling period depends on particular circuit conditions and details of thermistor design, but generally is several times the working time delay. In telephone relay circuits requiring a timing operation soon after previous use, the thermistor usually is connected so that it is short circuited by the relay contacts at the close of the working time delay interval. This permits the thermistor to cool during the period when the relay is locked up. If this period is sufficiently long, the thermistor is available for use as soon as the relay drops out. Time delay thermistors have been operated more than half a million times on life test with no significant change in their timing action.

To avoid the limitations of wide timing interval limits and extended cooling period between operations usually associated with the power driven time delay thermistor, a cooling time delay method of operation has been used. In this arrangement, two relays or the equivalent are employed and the thermistor is heated to a low resistance-value by passing a relatively large current through it for an interval short compared with the desired time interval. The current then is reduced automatically to a lower value and the thermistor cools until its resistance increases enough to reduce the current further and trip the working relay. This part of the operating cycle accounts for the greater part of the desired time interval. With this arrangement, the thermistor is available for re-use immediately after a completed timing interval, or, as a matter of fact, after any part of it. By proper choice of operating currents and circuit values, wide variations of voltage and ambient temperature may occur with relatively little effect upon the time interval. The principal variable left is the cooling time of the thermistor itself. This is fixed in a given thermistor unit, but may vary from unit to unit, depending upon dissipation constant and thermal capacity, as pointed out above.

In addition to their use as definite time delay devices, thermistors have been used in several related applications. Surges can be prevented from

operating relays or disturbing sensitive apparatus by introducing a thermistor in series with the circuit component which is to be protected. In case of a surge, the high initial resistance of the thermistor holds the surge current to a low value provided that the surge does not persist long enough to overcome the thermal inertia of the thermistor. The normal operating voltage, on the other hand, is applied long enough to lower the thermistor resistance to a negligible value, so that a normal operating current will flow after a short interval. In this way, the thermistor enables the circuit to distinguish between an undesired signal of short duration and a desired signal of longer duration even though the undesired impulse is several times higher in voltage than the signal.

Oscillators, Modulators and Amplifiers

A group of applications already explored in the laboratory but not put into engineering use includes oscillators, modulators and amplifiers for the low and audio-frequency range. If a thermistor is biased at a point on the negative slope portion of the steady-state volt-ampere characteristic, and if a small alternating voltage is then superposed on the direct voltage, a small alternating current will flow. If the thermistor has a small time constant, τ, and if the applied frequency is low enough, the alternating volt-ampere characteristic will follow the steady-state curve and dV/dI will be negative. As the frequency of the applied a-c voltage is increased, the value of the negative resistance decreases. At some critical frequency, f_c, the resistance is zero and the current is 90 degrees out of phase with the voltage. In the neighborhood of f_c, the thermistor acts like an inductance whose value is of the order of a henry. As the frequency is increased beyond f_c, the resistance is positive and increases steadily until it approaches the d-c value when the current and voltage are in phase. The critical frequency is given approximately by

$$f_c = 1/2\tau.$$

If τ can be made as small as 5×10^{-5} seconds, f_c is equal to 10,000 cycles per second and the thermistor would have an approximately constant negative resistance up to half this frequency. Point contact thermistors having such critical frequencies or even higher have been made in a number of laboratories. However, none of them have been made with sufficient reproducibility and constancy to be useful to the engineer. It has been shown both theoretically and experimentally that any negative resistance device can be used as an oscillator, a modulator, or an amplifier. With further development, it seems probable that thermistors will be used in these fields.

Summary

The general principles of thermistor operation and examples of specific uses have been given to facilitate a better understanding of them, with the feeling that such an understanding will be the basis for increased use of this new circuit and control element in technology.

References

1. Zur Elektrischen Leitfahigkeit von Kupferoxydul, W. P. Juse and B. W. Kurtschatow. Physikalische Zeitschrift Der Sowjetunion, Volume 2, 1932, pages 453–67.
2. Semi-conductors and Metals (book), A. H. Wilson. The University Press, Cambridge, England, 1939.
3. The Modern Theory of Solids (book), Frederick Seitz. McGraw-Hill Book Company, New York, N. Y., 1940.
4. Electronic Processes in Ionic Crystals (book), N. F. Mott and R. W. Gurney. The Clarendon Press, Oxford, England, 1940.
5. Die Elektronenleitfahigkeit von Festen Oxyden Verschiedener Valenzstufen, M. LeBlanc and H. Sachse. Physikalische Zeitschrift, Volume 32, 1931, pages 887–9.
6. Uber die Elektrizitatsleitung Anorganischer Stoffe mit Elektronenleitfahigkeit, Wilfried Meyer. Zeitschrift Fur Physik, Volume 85, 1933, pages 278–93.
7. Thermal Agitation of Electricity in Conductors, J. B. Johnson. Physical Review, Volume 32, July 1928, pages 97–113.
8. Spontaneous Resistance Fluctuations in Carbon Microphones and Other Granular Resistances, C. J. Christensen and G. L. Pearson. The Bell System Technical Journal, Volume 15, April 1936, pages 197–223.
9. Automatic Temperature Control for Aircraft, R. A. Gund. AIEE Transactions, Volume 64, 1945, October section, pages 730–34.
10. The Bridge Stabilized Oscillator, L. A. Meacham. Proc. IRE, Volume 26, October 1938, pages 1278–94.
11. Frequency Stabilized Oscillator, R. L. Shepherd and R. O. Wise. Proc. IRE, Volume 31, June 1943, pages 256–68.
12. A Pilot-Channel Regulator for the K-1 Carrier System, J. H. Bollman. Bell Laboratories Record, Volume 20, No. 10, June 1942, pages 258–62.
13. Thermistors, J. E. Tweeddale. Western Electric Oscillator, December 1945, pages 3–5, 34–7.
14. Thermistor Technics, J. C. Johnson. Electronic Industries, Volume 4, August 1945, pages 74–7.
15. Volume Limiter for Leased-Line Service, J. A. Weller. Bell Laboratories Record, Volume 23, No. 3, March 1945, pages 72–5.

Abstracts of Technical Articles by Bell System Authors

Capacitors—Their Use in Electronic Circuits.[1] M. Brotherton. This book tells how to choose and use capacitors for electronic circuits. It explains the basic factors which control the characteristics of capacitors and determine their proper operation. It helps to provide that broad understanding of the capacitor problem which is indispensable to the efficient design of circuits. It tells the circuit designer what he must understand and consider in transforming capacitance from a circuit symbol into a practical item of apparatus capable of meeting the growing severity of today's operation requirements.

Mica Capacitors for Carrier Telephone Systems.[2] A. J. Christopher and J. A. Kater. Silvered mica capacitors, because of their inherently high capacitance stability with temperature changes and with age, now are used widely in oscillators, networks, and other frequency determining circuits in the Bell Telephone System. Their use in place of the previous dry stack type, consisting of alternate layers of mica and foil clamped under high pressures, has made possible considerable manufacturing economies in addition to improving the transmission performance of carrier telephone circuits. These economies are the result of their relatively simple unit construction and the ease of adjustment to the very close capacitance tolerance required.

Visible Speech Translators with External Phosphors.[3] Homer Dudley and Otto O. Gruenz, Jr. This paper describes some experimental apparatus built to give a passing display of visible speech patterns. These patterns show the analysis of speech on an intensity-frequency-time basis and move past the reader like a printed line. The apparatus has been called a translator as it converts speech intended for aural perception into a form suitable for visual prception. The phosphor employed is not in a cathode-ray tube but in the open on a belt or drum.

The Pitch, Loudness and Quality of Musical Tones (A demonstration-lecture introducing the new Tone Synthesizer).[4] Harvey Fletcher. Relations are given in this paper which show how the pitch of a musical tone

[1] Published by D. Van Nostrand Company, Inc., New York, N. Y., 1946.
[2] *Elec. Engg., Transactions Section,* October 1946.
[3] *Jour. Acous. Soc. Amer.,* July 1946.
[4] *Amer. Jour. of Physics,* July–August 1946.

depends upon the frequency, the intensity and the overtone structure of the sound wave transmitting the tone. Similar relations are also given which show how the loudness and the quality depend upon these same three physical characteristics of the sound wave. These relationships were demonstrated by using the new Tone Synthesizer. By means of this instrument one is able to imitate the quality, pitch and intensity of any musical tone and also to produce many combinations which are not now used in music.

The Sound Spectrograph.[5] W. Koenig, H. K. Dunn, and L. Y. Lacy. The sound spectrograph is a wave analyzer which produces a permanent visual record showing the distribution of energy in both frequency and time. This paper describes the operation of this device, and shows the mechanical arrangements and the electrical circuits in a particular model. Some of the problems encountered in this type of analysis are discussed, particularly those arising from the necessity for handling and portraying a wide range of component levels in a complex wave such as speech. Spectrograms are shown for a wide variety of sounds, including voice sounds, animal and bird sounds, music, frequency modulations, and miscellaneous familiar sounds.

Geometrical Characterizations of Some Families of Dynamical Trajectories.[6] L. A. MacColl. A broad problem in differential geometry is that of characterizing, by a set of geometrical properties, the family of curves which is defined by a given system of differential equations, of a more or less special form. The problem has been studied especially by Kasner and his students, and characterizations have been obtained for various families of curves which are of geometrical or physical importance. However, the interesting problem of characterizing the family of trajectories of an electrified particle moving in a static magnetic field does not seem to have been considered heretofore. The present paper gives the principal results of a study of this problem.

Visible Speech Cathode-Ray Translator.[7] R. R. Riesz and L. Schott. A system has been developed whereby speech analysis patterns are made continuously visible on the moving luminescent screen of a special cathode-ray tube. The screen is a cylindrical band that rotates with the tube about a vertical axis. The electron beam always excites the screen in the same vertical plane. Because of the persistence of the screen phosphor and the rotation of the tube, the impressed patterns are spread out along a horizon-

[5] *Jour. Acous. Soc. Amer.*, July 1946.
[6] *Amer. Math. Soc. Transactions*, July 1946.
[7] *Jour. Acous. Soc. Amer.*, July 1946.

tal time axis so that speech over an interval of a second or more is always visible. The upper portion of the screen portrays a spectrum analysis and the lower portion a pitch analysis of the speech sounds. The frequency band up to 3500 cycles is divided into 12 contiguous sub-bands by filters. The average speech energy in the sub-bands is scanned and made to control the excitation of the screen by the electron beam which is swept synchronously across the screen in the vertical direction. A pitch analyzer produces a d-c. voltage proportional to the instantaneous fundamental frequency of the speech and this controls the width of a band of luminescence that the electron beam produces in the lower part of the screen. The translator had been used in a training program to study the readability of visible speech patterns.

Derivatives of Composite Functions.[8] JOHN RIORDAN. The object of this note is to show the relation of the Y polynomials of E. T. Bell, first to the formula of DiBruno for the nth derivative of a function of a function, then to the more general case of a function of many functions. The subject belongs to the algebra of analysis in the sense of Menger; all that is asked is the relation of the derivative of the composite function to the derivatives of its component functions when they exist and no questions of analysis are examined.

The Portrayal of Visible Speech.[9] J. C. STEINBERG AND N. R. FRENCH. This paper discusses the objectives and requirements in the protrayal of visible patterns of speech from the viewpoint of their effects on the legibility of the patterns. The portrayal involves an intensity-frequency-time analysis of speech and the display of the results of the analysis to the eye. Procedures for accomplishing this are discussed in relation to information on the reading of print and on the characteristics of speech and its interpretation by the ear. Also methods of evaluating the legibility of the visible patterns are described.

Short Survey of Japanese Radar—I.[10] ROGER I. WILKINSON. The result of a study made immediately following the fall of Japan and recently made available for public information, this two-part report is designed to present a quick over-all evaluation of Japanese radar, its history and development. As the Japanese army and navy developed their radar equipment independently of each other, Part I of this article concentrates on the army's contributions.

[8] *Amer. Math. Soc. Bulletin*, August 1946.
[9] *Jour. Acous. Soc. Amer.*, July 1946.
[10] *Elec. Engg.*, Aug.–Sept. 1946.

A Variation on the Gain Formula for Feedback Amlifiers for a Certain Driving-Impedance Configuration.[11] T. W. WINTERNITZ. An expression for the gain of a feedback amplifier, in which the source impedance is the only significant impedance across which the feedback voltage is developed, is derived. As examples of the use of this expression, it is then applied to three common circuits in order to obtain their response to a Heaviside unit step-voltage input.

[11] *Proc. I.R.E.*, September 1946.

Contributors to This Issue

Joseph A. Becker, A.B., Cornell University 1918; PhD., Cornell University, 1922. National Research Fellow, California Institute of Technology, 1922–24; Asst. Prof. of Physics, Stanford University, 1924. Engineering Dept., Western Electric Company, 1924–1925; Bell Telephone Laboratores, 1925–. Mr. Becker has worked in the fields of X-Rays, magnetism, thermionic emission and adsorption, particularly in oxide coated filaments, the properties of semiconductors, as applied in varistors and thermistors.

W. R. Bennett, B. S., Oregon State College, 1925; A.M., Columbia University, 1928. Bell Telephone Laboratories, 1925–. Mr. Bennett has been active in the design and testing of multichannel communication systems, particularly with regard to modulation processes and the effects of nonlinear distortion. As a member of the Transmission Research Department, he is now engaged in the study of pulse modulation techniques for sending telephone channels by microwave radio relay.

C. B. Green, Ohio State University, B.A. 1927; M.A. in Physics, 1928. Additional graduate work at Columbia University. Bell Telephone Laboratories, 1928–. For ten years Mr. Green was concerned with transmission development for telephotography and television systems and with the design of vacuum tubes. Since 1938 he has been engaged in the development and application of thermistors.

J. P. Kinzer, M. E., Stevens Institute of Technology, 1925. B.C.E., Brooklyn Polytechnic Institute, 1933. Bell Telephone Laboratories, 1925–. Mr. Kinzer's work has been in the development of carrier telephone repeaters; during the war his attention was directed to investigation of the mathematical problems involved in cavity resonators.

W. P. Mason, B.S. in E.E., Univ. of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged principally in investigating the properties and applications of piezoelectric crystals and in the study of ultrasonics.

R. S. Ohl, B. S. in Electro-Chemical Engineering, Pennsylvania State College, 1918; U. S. Army, 1918 (2nd Lieutenant, Signal Corps); Vacuum tube development, Westinghouse Lamp Company, 1919–21; Instructor in

Physics, University of Colorado, 1921–1922. Department of Development and Research, American Telephone and Telegraph Company, 1922–27; Bell Telephone Laboratories, 1927–. Mr. Ohl has been engaged in various exploratory phases of radio research, the results of which have led to numerous patents. For the past ten or more years he has been working on some of the problems encountered in the use of millimeter radio waves.

G. L. Pearson, A. B., Willamette University, 1926; M. A. in Physics, Stanford University, 1929. Bell Telephone Laboratories, 1929–. Mr. Pearson is in the Physical Research Department where he has been engaged in the study of noise in electric circuits and the properties of electronic semiconductors.

J. H. Scaff, B.S.E. in Chemical Engineering, University of Michigan, 1929. Bell Telephone Laboratories, 1929–. Mr. Scaff's early work in the Laboratories was concerned with metallurgical investigations of impurities in metals with particular reference to soft magnetic materials. During the war he was project engineer for the development of silicon and germanium crystal rectifiers for radar applications. At the present time, he is responsible for metallurgical work on varistor and magnetic materials.

I. G. Wilson, B.S. and M.E., University of Kentucky, 1921. Western Electric Co., EngineeringDepartment, 1921–25. Bell Telephone Laboratories, 1925–. Mr. Wilson has been engaged in the development of amplifiers for broad-band systems. During the war he was project engineer in charge of the design of resonant cavities for radar testing.

VOLUME XXVI APRIL, 1947 NO. 2

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Public Library
Kansas City, Mo

Radar Antennas.....*H. T. Friis and W. D. Lewis* 219

Probability Functions for the Modulus and Angle of the Normal Complex Variate.....*Ray S. Hoyt* 318

Spectrum Analysis of Pulse Modulated Waves *J. C. Lozier* 360

Abstracts of Technical Articles by Bell System Authors.. 388

Contributors to This Issue..... 394

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50¢ per copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORS

R. W. King J. O. Perrine

EDITORIAL BOARD

W. H. Harrison
O. B. Blackwell
H. S. Osborne
J. J. Pilliod
O. E. Buckley
M. J. Kelly
A. B. Clark
S. Bracken

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are 50 cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1947
American Telephone and Telegraph Company

PRINTED IN U. S. A.

BOOKS AND PERIODICALS FOR WORLD RECOVERY

The desperate and continued need for American publications to serve as tools of physical and intellectual reconstruction abroad has been made vividly apparent by appeals from scholars in many lands. The American Book Center for War Devastated Libraries has been urged to continue meeting this need at least through 1947. The Book Center is therefore making a renewed appeal for American books and periodicals—*for technical and scholarly books and periodicals in all fields* and particularly for *publications of the past ten years.* We shall particularly welcome complete or incomplete files of the *Bell System Technical Journal.*

The generous support which has been given to the Book Center has made it possible to ship more than 700,000 volumes abroad in the past year. It is hoped to double this amount before the Book Center closes. The books and periodicals which individuals as well as institutional libraries can spare are urgently needed and will help in the reconstruction which must preface world understanding and peace.

Ship your contributions to the American Book Center, c/o The Library of Congress, Washington 25, D. C., freight prepaid, or write to the Center for further information.

The Bell System Technical Journal

Vol. XXVI April, 1947 No. 2

Radar Antennas

By H. T. FRIIS and W. D. LEWIS

Table of Contents

Introduction ... 220
Part I—Electrical Principles ... 224
1. General ... 224
2. Transmission Principles ... 226
2.1 Gain and Effective Area of an Antenna ... 226
Definition of Gain ... 226
Definition of Effective Area ... 226
2.2 Relationship between Gain and Effective Area ... 227
2.3 The Ratio G/A for a Small Current Element ... 227
2.4 The General Transmission Formula ... 230
2.5 The Reradiation Formula ... 230
2.6 The Plane, Linearly Polarized Electromagnetic Wave ... 231
3. Wave Front Analysis ... 232
3.1 The Huygens Source ... 233
3.2 Gain and Effective Area of an Ideal Antenna ... 235
3.3 Gain and Effective Area of an Antenna with Aperture in a Plane and with Arbitrary Phase and Amplitude ... 236
3.4 The Significance of the Pattern of a Radar Antenna ... 237
3.5 Pattern in Terms of Antenna Wave Front ... 238
3.6 Pattern of an Ideal Rectangular Antenna ... 239
3.7 Effect on Pattern of Amplitude Taper ... 240
3.8 Effect on Pattern of Linear Phase Variation ... 241
3.9 Effect on Pattern of Square Law Phase Variation ... 242
3.10 Effect on Pattern of Cubic Phase Variation ... 244
3.11 Two General Methods ... 245
3.12 Arrays ... 246
3.13 Limitations to Wave Front Theory ... 246
4. Application of General Principles ... 247
Part II—Methods of Antenna Construction ... 247
5. General ... 247
6. Classification of Methods ... 248
7. Basic Design Formulation ... 250
7.1 Dimensions of the Aperture ... 250
7.2 Amplitude Distribution ... 251
7.3 Phase Control ... 251
8. Parabolic Antennas ... 251
8.1 Control of Phase ... 251
8.2 Control of Amplitude ... 253
8.3 Choice of Configuration ... 254
8.4 Feeds for Paraboloids ... 258
8.5 Parabolic Cylinders between Parallel Plates ... 260
8.6 Line Sources for Parabolic Cylinders ... 262
8.7 Tolerances in Parabolic Antennas ... 264
9. Metal Plate Lenses ... 266
9.1 Lens Antenna Configurations ... 269
9.2 Tolerances in Metal Plate Lenses ... 269
9.3 Advantages of Metal Plate Lenses ... 270
10. Cosecant Antennas ... 270
10.1 Cosecant Antennas based on the Paraboloid ... 271
10.2 Cylindrical Cosecant Antennas ... 274

11. Lobing 274
11.1 Lobe Switching 275
11.2 Conical Lobing 276
12. Rapid Scanning 276
12.1 Mechanical Scanning 277
12.2 Array Scanning 278
12.3 Optical Scanning 282
Part III—Military Radar Antennas Developed by the Bell Laboratories 284
13. General 284
14. Naval Shipborne Radar Antennas 286
14.1 The SE Antenna 286
14.2 The SL Antenna 286
14.3 The SJ Submarine Radar Antenna 291
14.4 The Modified SJ/Mark 27 Radar Antenna 294
14.5 The SH and Mark 16 Radar Antennas 294
14.6 Antennas for Early Fire Control Radars 297
14.7 A Shipborne Anti-Aircraft Fire Control Antenna 298
14.8 The Polyrod Fire Control Antenna 300
14.9 The Rocking Horse Fire Control Antenna 301
14.10 The Mark 19 Radar Antenna 302
14.11 The Mark 28 Radar Antenna 305
14.12 A 3 cm Anti-Aircraft Radar Antenna 307
15. Land Based Radar Antennas 307
15.1 The SCR-545 Radar "Search" and "Track" Antennas 307
15.2 The AN/TPS-1A Portable Search Antenna 309
16. Airborne Radar Antennas 312
16.1 The AN/APS-4 Antenna 312
16.2 The SCR-520, SCR-717 and SCR-720 Antennas 313
16.3 The AN/APQ-7 Radar Bombsight Antenna 315

Introduction

RADAR proved to be one of the most important technical achievements of World War II. It has many sources, some as far back as the nineteenth century, yet its rapid wartime growth was the result of military necessity. This development will continue, for radar has increasing applications in a peacetime world.

In this paper we will discuss an indispensable part of radar—the antenna. In a radar system the antenna function is two-fold. It both projects into space each transmitted radar pulse, and collects from space each received reflected signal. Usually but not always a single antenna performs both functions.

The effectiveness of a radar is influenced decisively by the nature and quality of its antenna. The greatest range at which the radar can detect a target, the accuracy with which the direction to the target can be determined and the degree with which the target can be discriminated from its background or other targets all depend to a large extent on electrical properties of the antenna. The angular sector which the antenna can mechanically or electrically scan is the sector from which the radar can provide information. The scanning rate determines the frequency with which a tactical or navigational situation can be examined.

Radar antennas are as numerous in kind as radars. The unique character and particular functions of a radar are often most clearly evident in the design of its antenna. Antennas must be designed for viewing planes from the ground, the ground from planes and planes from other planes. They must see ships from the shore, from the air, from other ships, and from submarines. In modern warfare any tactical situation may require one or several radars and each radar must have one or more antennas.

Radar waves are almost exclusively in the centimeter or microwave region, yet even the basic microwave techniques are relatively new to the radio art. Radar demanded antenna gains and directivities far greater than those previously employed. Special military situations required antennas with beam shapes and scanning characteristics never imagined by communication engineers.

It is natural that war should have turned our efforts so strongly in the direction of radar. But that these efforts were so richly and quickly rewarded was due in large part to the firm technical foundations that had been laid in the period immediately preceeding the war. When, for the common good, all privately held technical information was poured into one pool, all ingredients of radar, and of radar antennas in particular, were found to be present.

A significant contribution of the Bell System to this fund of technical knowledge was its familiarity with microwave techniques. Though Hertz himself had performed radio experiments in the present microwave region, continuous wave techniques remained for decades at longer wavelengths. However, because of its interest in new communication channels and broader bands the Bell System has throughout the past thirty years vigorously pushed continuous wave techniques toward the direction of shorter waves. By the middle nineteen-thirties members of the Radio Research Department of the Bell Laboratories were working within the centimeter region.

Several aspects of this research and development appear now as particularly important. In the first place it is obvious that knowledge of how to generate and transmit microwaves is an essential factor in radar. Many lower frequency oscillator and transmission line techniques are inapplicable in the microwave region. The Bell Laboratories has been constantly concerned with the development of generators which would work at higher and higher frequencies. Its broad familiarity with coaxial cable problems and in particular its pioneering work with waveguides provided the answers to many radar antenna problems.

Another telling factor was the emphasis placed upon measurement. Only through measurements can the planners and designers of equip-

ment hope to evaluate performance, to chose between alternatives or to see the directions of improvement. Measuring techniques employing double detection receivers and intermediate frequency amplifiers had long been in use at the Holmdel Radio Laboratory. By employing these techniques radar engineers were able to make more sensitive and accurate measurements than would have been possible with single detection.

Antennas are as old as radio. Radar antennas though different in form are identical in principle with those used by Hertz and Marconi. Consequently experience with communication antennas provided a valuable background for radar antenna design. As an example of the importance of this background it can be recalled that a series of experi-

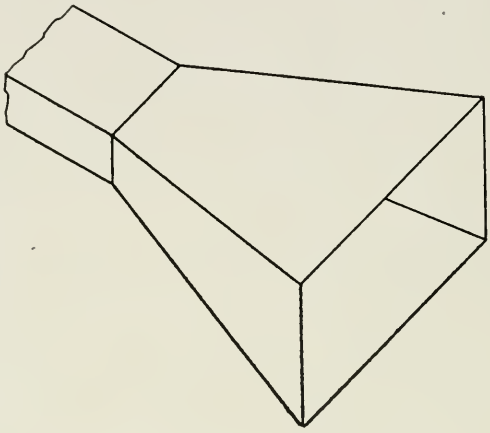


Fig. 1—An Electromagnetic Horn.

ments with short wave antennas for Transatlantic radio telephone service had culminated in 1936 in a scanning array of rhombic antennas. The essential principles of this array were later applied to shipborne fire control antenna which was remarkable and valuable because of the early date at which it incorporated modern rapid scanning features.

In addition to the antenna arts which arose directly out of communication problems at lower frequencies some research specifically on microwave antennas was under way before the war. Early workers in waveguides noticed that an open ended waveguide will radiate directly into space. It is not surprising therefore that these workers developed the electromagnetic horn, which is essentially a waveguide tapered out to an aperture (Fig. 1).

One of the first used and simplest radio antennas is the dipole (Fig.

2). Current oscillating in the dipole generates electromagnetic waves which travel out with the velocity of light. A single dipole is fairly non-directive and consequently produces a relatively weak field at a distance. When the wave-length is short the field of a dipole in a

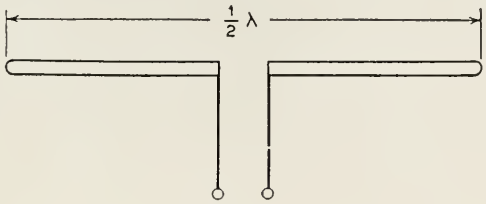


Fig. 2—A Microwave Dipole.

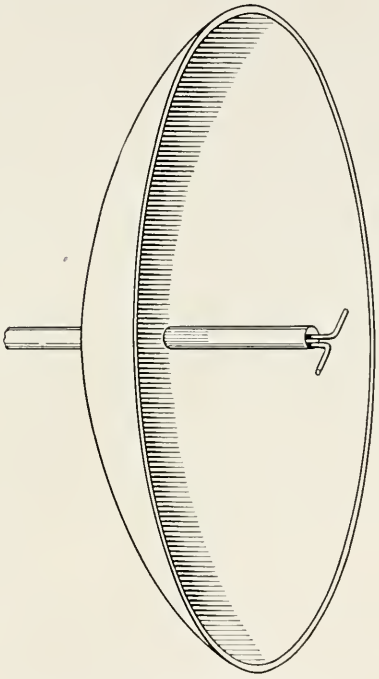


Fig. 3—A Dipole Fed Paraboloid.

chosen direction can be increased many times by introducing a reflector which directs or 'focusses' the energy.

In communication antennas the focussing reflector is most commonly a reflecting wire array. Even at an early date in radar the wavelength was so short that 'optical' reflectors could be used. These were

sometimes paraboloids similar to those used in searchlights (Fig. 3). Sometimes they were parabolic cylinders as in the Mark III, an early shipborne fire control radar developed at the Whippany Radio Laboratory.

From these relatively simple roots, the communication antenna, the electromagnetic horn and the optical reflector, radar antennas were developed tremendously during the war. That this development in the Bell Laboratories was so well able to meet demands placed on it was due in large part to the solid foundation of experience possessed by the Research and Development groups of the Laboratories. Free interchange of individuals and information between the Laboratories and other groups, both in the United States and Great Britain, also contributed greatly to the success of radar antenna development.

Because of its accelerated wartime expansion the present radar antenna field is immense. It is still growing. It would be impossible for any single individual or group to master all details of this field, yet its broad outline can be grasped without difficulty.

The purpose of this paper is two-fold, both to provide a general discussion of radar antennas and to summarize the results of radar antenna research and development at the Bell Laboratories. Part I is a discussion of the basic electrical principles which concern radar antennas. In Part II we will outline the most common methods of radar antenna construction. Practical military antennas developed by the Bell Laboratories will be described in Part III.

The reader who is interested in general familiarity with the over all result rather than with technical features of design may proceed directly from this part to Part III.

PART I

ELECTRICAL PRINCIPLES

1. General

Radar antenna design depends basically on the same broad principles which underlie any other engineering design. The radar antenna designer can afford to neglect no aspect of his problem which has a bearing on the final product. Mechanical, chemical, and manufacturing considerations are among those which must be taken into account.

It is the electrical character of the antenna, however, which is connected most directly with the radar performance. In addition it is through attention to the electrical design problems that the greatest number of novel antennas have been introduced and it is from the electrical viewpoint that the new techniques can best be understood.

An antenna is an electromagnetic device and as such can be understood

through the application of electromagnetic theory. Maxwell's equations provide a general and accurate foundation for antenna theory. They are the governing authority to which the antenna designer may refer directly when problems of a fundamental or baffling nature must be solved.

It is usually impracticable to obtain theoretically exact and simple solutions to useful antenna problems by applying Maxwell's Equations directly. We can, however, use them to derive simpler useful theories. These theories provide us with powerful analytical tools.

Lumped circuit theory is a tool of this sort which is of immense practical importance to electrical and radio engineers. As the frequency becomes higher the approximations on which lumped circuit theory is based become inaccurate and engineers find that they must consider distributed inductances and capacitances. The realm of transmission line theory has been invaded.

Transmission line theory is of the utmost importance in radar antenna design. In the first place the microwave energy must be brought to the antenna terminals over a transmission line. This feed line is usually a coaxial or a wave-guide. It must not break down under the voltage which accompanies a transmitted pulse. It must be as nearly lossless and reflectionless as possible and it must be matched properly to the antenna terminals.

The importance of a good understanding of transmission line theory does not end at the antenna terminals. In any antenna the energy to be transmitted must be distributed in the antenna structure in such a way that the desired radiation characteristics will be obtained. This may be done with transmission lines, in which case the importance of transmission line theory is obvious. It may be done by 'optical' methods. If so, certain transmission line concepts and methods will still be useful.

While it is true that transmission line theory is important it is not necessary to give a treatment of it in this paper. Adequate theoretical discussions can be found elsewhere in several sources.[1] It is enough at this point to indicate the need for a practical understanding of transmission line principles, a need which will be particularly evident in Part II, Methods of Antenna Construction.

We may, if we like, think of the whole radar transmission problem in terms of transmission line theory. The antenna then appears as a transformer between the feed line and transmission modes in free space. We cannot, however, apply this picture to details with much effectiveness unless we have some understanding of radiation.

In the sections to follow we shall deal with some theoretical aspects of radiation. We shall begin with a discussion of fundamental *transmission*

[1] See, for example, S. A. Schelkunoff, *Electromagnetic Waves*, D. Van Nostrand Co., Inc., 1943, in particular, Chapters VII and VIII, or F. E. Terman, *Radio Engineer's Handbook*, McGraw-Hill Book Co., Inc., 1943, Section 3.

principles. This discussion is applicable to all antennas regardless of how they are made or used. When applied to radar antennas it deals chiefly with those properties of the antenna which affect the radar range.

Almost all microwave radar antennas are large when measured in wavelengths. When used as transmitting antennas they produce desired radiation characteristics by distributing the transmitted energy over an area or 'wave front'. The relationships between the phase and amplitude of electrical intensity in this wave front and the radiation characteristics of the antenna are predicted by *wave front analysis.* Wave front analysis is essentially the optical theory of diffraction. Although approximate it applies excellently to the majority of radar antenna radiation problems. We shall discuss wave front analysis in Section 3.

2. Transmission Principles

2.1 *Gain and Effective Area of an Antenna*

An extremely important property of any radar antenna is its ability to project a signal to a distant target. The *gain* of the antenna is a number which provides a quantitative measure of this ability. Another important property of a radar antenna is its ability to collect reflected power which is returning from a distant target. The *effective area* of the antenna is a quantitative measure of this ability. In this section these two quantities will be defined, and a simple relation between them will be derived. Their importance to radar range will be established.

Definition of Gain. When power is fed into the terminals of an antenna some of it will be lost in heat and some will be radiated. The *gain* G of the antenna can be defined as the ratio

$$G = P/P_0 \tag{1}$$

where P is the power flow per unit area in the plane linearly polarized electromagnetic wave which the antenna causes in a distant region usually in the direction of maximum radiation and P_0 is the power flow per unit area which would have been produced if all the power fed into the terminals had been radiated equally in all directions in space.

Definition of Effective Area. When a plane linearly polarized electromagnetic wave is incident on the receiving antenna, received power P_R will be available at the terminals of the antenna. The *effective area* of the antenna is defined, by the equation

$$A = P_R/P' \tag{2}$$

where P' is the power per unit area in the incident wave. In other words the received power is equal to the power flow through an area that is equal to the effective area of the antenna.

2.2 *Relationship between Gain and Effective Area*

Figure 4 shows a radio circuit in free space made up of a transmitting antenna T and a receiving antenna R. If the transmitted power P_T had

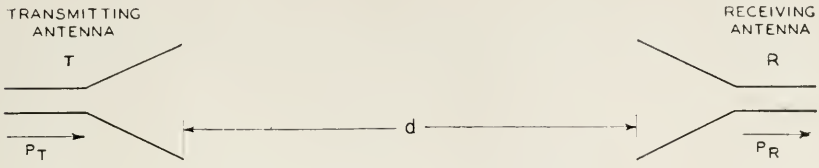


Fig. 4—Radio Circuit in Free Space.

been radiated equally in all directions, the power flow per unit area at the receiving antenna would be

$$P_0 = P_T \frac{1}{4\pi d^2} \tag{3}$$

Definition (1) gives, therefore, for the power flow per unit area at the receiving antenna

$$P = P_0 G_T = \frac{P_T G_T}{4\pi d^2} \tag{4}$$

and definition (2) gives for the received power

$$P_R = P A_R = \frac{P_T G_T A_R}{4\pi d^2} \tag{5}$$

From the law of reciprocity it follows that the same power is transferred if the transmitting and receiving roles are reversed. By (5) it is thus evident that

$$G_T A_R = G_R A_T$$

or

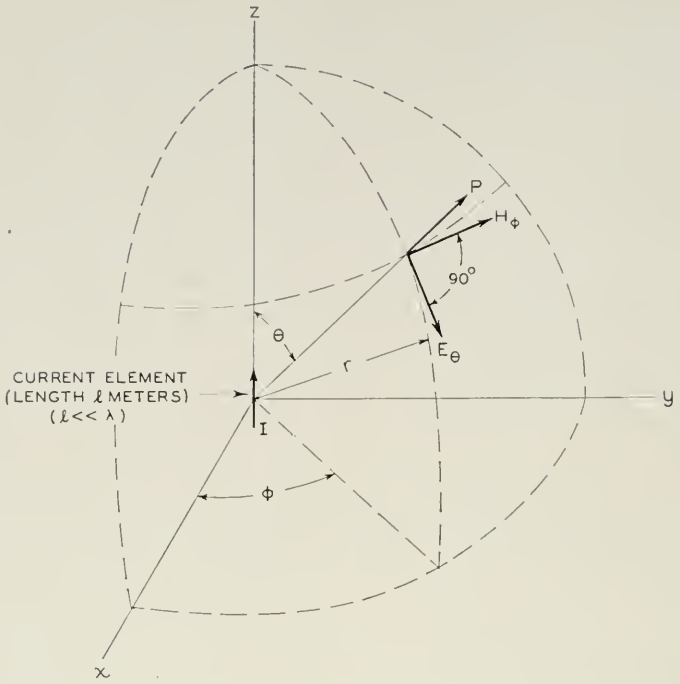
$$G_T/A_T = G_R/A_R \tag{6}$$

Equation (6) shows that the ratio of the gain and effective area has the same constant value for all antennas at a given frequency. It is necessary, therefore, to calculate this ratio only for a simple and well known antenna such as a small dipole or uniform current element.

2.3 *The Ratio G/A for a Small Current Element*

In Fig. 5 are given formulas[2] in M.K.S. units for the free space radiation from a small current element with no heat loss. We have assumed that

[2] See S. A. Schelkunoff, *Electromagnetic Waves*, D. Van Nostrand Co., Inc., 1943, p. 133



FOR $r >> \lambda$:

$$\text{MAGNETIC INTENSITY} = H_\phi = i\frac{I\ell}{2\lambda r}e^{-i\frac{2\pi}{\lambda}r}\sin\theta \quad \frac{\text{AMPERES}}{\text{METER}} \tag{1}$$

$$\text{ELECTRIC INTENSITY} = E_\theta = 120\pi H_\phi \quad \frac{\text{VOLTS}}{\text{METER}} \tag{2}$$

$$\text{POWER FLOW} = P = |H_\phi E_\theta| = 30\pi\left[\frac{I\ell}{\lambda r}\right]^2 \sin^2\theta \quad \frac{\text{WATTS}}{\text{METER}^2} \tag{3}$$

$$P \text{ IS MAXIMUM FOR } \theta = 90^\circ \text{ i.e., } P_{MAX} = 30\pi\left[\frac{I\ell}{\lambda r}\right]^2 \frac{\text{WATTS}}{\text{METER}^2} \tag{4}$$

POWER FLOW ACROSS SPHERE OF RADIUS r OR

$$\text{TOTAL RADIATION} = W = \int_0^\pi P\, 2\pi r \sin\theta\, r\,d\theta = 80\pi^2\left[\frac{I\ell}{\lambda}\right]^2 \text{WATTS} \tag{5}$$

$$\text{RADIATION RESISTANCE} = R_{RAD} = \frac{W}{I^2} = 80\pi^2\left(\frac{\ell}{\lambda}\right)^2 \text{OHMS} \tag{6}$$

$$\text{BY (4) AND (5)}: P_{MAX} = \frac{3}{8\pi r^2} W \quad \frac{\text{WATTS}}{\text{METER}^2} \tag{7}$$

Fig. 5—Free Space Radiation from a Small Current Element with Uniform Current I Amperes over its Entire Length.

this element is centered at the origin of a rectangular coordinate system and that it lies along the Z axis. At a large distance r from the element

the maximum power flow per unit area occurs in a direction normal to it and is given by

$$P_{\max} = \frac{3W}{8\pi r^2} \frac{\text{watts}}{\text{meter}^2} \tag{7}$$

where W is the total radiated power. If W had been radiated equally in all directions the power flow per unit area would be

$$P_0 = \frac{W}{4\pi r^2} \frac{\text{watts}}{\text{meters}^2} \tag{8}$$

It follows that the gain of the small current element is

$$G_{\text{dipole}} = \frac{P_{\max}}{P_0} = 1.5 \tag{9}$$

The effective area of the dipole will now be calculated. When it is used to receive a plane linearly polarized electromagnetic wave, the available output power is equal to the induced voltage squared divided by four times the radiation resistance. Thus

$$P_R = \frac{E^2 \ell^2}{4R_{\text{rad}}} \text{ Watts} \tag{10}$$

where E is the effective value of the electric field of the wave, ℓ is the length of the current element and R_{rad} is the radiation resistance of the current element. From Fig. 5 we see that $R_{\text{rad}} = \frac{80\pi^2\ell^2}{\lambda^2}$ ohms. Since the power flow per unit area is equal to the electric field squared divided by the impedance of free space, in other words $P_0 = \frac{E^2}{120\pi}$ we have

$$A_{\text{dipole}} = \frac{P_R}{P_0} = \frac{3\lambda^2}{8\pi} \text{ meter}^2 \tag{11}$$

We combine formulas (9) and (11) to find that

$$\frac{G_{\text{dipole}}}{A_{\text{dipole}}} = \frac{4\pi}{\lambda^2}$$

Since, as proved in 2.2 this ratio is the same for all antennas, it follows that for any antenna

$$\frac{G}{A} = \frac{4\pi}{\lambda^2} \tag{12}$$

2.4 *The General Transmission Formula*

Transmission loss between transmitter and receiver through the radio circuit shown in Fig. 4 was given by equation (5). By substituting the relation (12) into (5) we can obtain the simple free space transmission formula:

$$P_R = P_T \frac{A_T A_R}{\lambda^2 d^2} \text{ watts} \tag{13}$$

Although this formula applies to free space only it is believed to be as useful in radio engineering as Ohm's law is in circuit engineering.

2.5 *The Reradiation Formula*

One further relation, the radar reflection formula is of particular interest. Consider the situation illustrated in Fig. 6. Let P_T be the power radiated

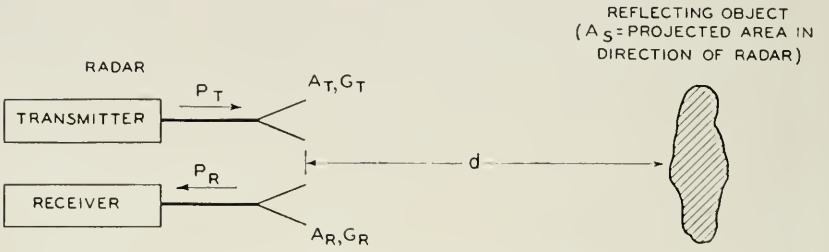


Fig. 6—Radar with Separate Receiving and Transmitting Antennas.

from an antenna with effective area A_T, A_S the area of a reflecting object at distance d from the antenna and P_R the power received by an antenna of effective area A_R. By equation (13) the power striking A_S is $\frac{P_T A_T A_S}{\lambda^2 d^2}$. If this power were reradiated equally in all directions the reflected power flow at the receiving antenna would be $\frac{P_T A_T A_S}{4\pi d^4 \lambda^2}$ but since the average reradiation is larger toward the receiving antenna, the power flow per unit area there is usually $K \frac{P_T A_T A_S}{4\pi d^4 \lambda^2}$ where $K > 1$. It follows from (2) that

$$P_R = K \frac{P_T A_T A_R A_S}{4\pi \lambda^2 d^4} \tag{14}$$

Formula (14) shows clearly why the use of large and efficient antennas will greatly increase the radar range.

Formula (14) applies to free space only. Application to other conditions

may require corrections for the effect of the "ground", and for the effect of the transmission medium, which are beyond the scope of this paper.

2.6 *The Plane, Linearly Polarized Electromagnetic Wave*

In the foregoing sections we have referred several times to 'plane, linearly polarized electromagnetic waves'. These waves occur so commonly in antenna theory and practice that it is worth while to discuss them further here.

Some properties of linearly polarized, plane electromagnetic waves are illustrated in Fig. 7. At any point in the wave there is an electric field and a magnetic field. These fields are vectorial in nature and are at right angles to each other and to the direction of propagation. It is customary to give the magnitude of the electric field only.

If we use the M.K.S. system of units the magnitudes of the fields are expressed in familiar units. Electric intensity appears as volts per meter and magnetic intensity as amperes per meter. The ratio of electric to magnetic intensity has a value of 120π or about 377 ohms. This is the 'impedance' of free space. The power flow per unit area is expressed in watts per square meter. We see, therefore, that the electromagnetic wave is a means for carrying energy not entirely unlike a familiar two wire line or a coaxial cable.

Electromagnetic waves are generated when oscillating currents flow in conductors. We could generate a plane linearly polarized electromagnetic wave with a uniphase current sheet consisting of a network of fine wires backed up with a conducting reflector as shown in Fig. 7. This wave could be absorbed by a plane resistance sheet with a resistivity of 377 ohms, also backed up by a conducting sheet. The perfectly conducting reflecting sheets put infinite impedances in parallel with the current sheet and the resistance sheet, since each of these reflecting sheets has a zero impedance at a spacing of a quarter wavelength.

A perfectly plane electromagnetic wave can exist only under certain ideal conditions. It must be either infinite in extent or bounded appropriately by perfect electric and magnetic conductors. Nevertheless thinking in terms of plane electromagnetic waves is common and extremely useful. In the first place the waves produced over a small region at a great distance from any radiator are essentially plane. Arguments concerning receiving antennas therefore generally assume that the incident waves are plane. In the second place an antenna which has dimensions of many wavelengths can be analyzed with considerable profit on the basis of the assumption that it transmits by producing a nearly plane electromagnetic wave across its aperture. This method of analysis can be applied to the majority of microwave radar antennas, and will be discussed in the following sections.

3. Wave Front Analysis

The fundamental design question is "How to get what we want?" In a radar antenna we want specified radiation characteristics; gain, pattern and polarization. Electromagnetic theory tells us that if all electric and magnetic currents in an antenna are known its radiation characteristics may be derived with the help of Maxwell's Equations. However, the essence of electromagnetic theory insofar as it is of use to the radar antenna

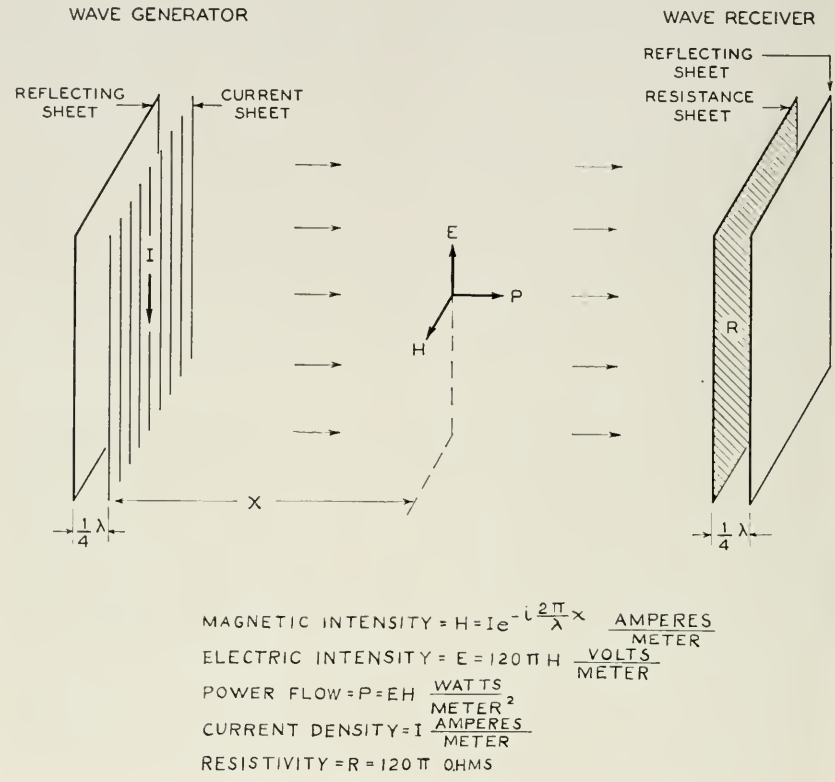


Fig. 7—Linearly Polarized Plane Electromagnetic Waves.

designer can usually be expressed in a simpler, more easily visualized and thus more useful form. This simpler method we call *wave front analysis*.

In a transmitting microwave antenna the power to be radiated is used to produce currents in antenna elements which are distributed in space. This distribution is usually over an area, it may be discrete as with a dipole array or it may be continuous as in an electromagnetic horn or paraboloid. These currents generate an advancing electromagnetic wave over the aperture of

the antenna. The amplitude, phase and polarization of the electric intensity in portions of the wave are determined by the currents in the antenna and thus by the details of the antenna structure. This advancing wave can be called the 'wave front' of the antenna.

When the wave front of an antenna is known its radiation characteristics may be calculated. Each portion of the wave front can be regarded as a secondary or 'Huygens' source of known electric intensity, phase and polarization. At any other point in space the electric intensity, phase and polarization due to a Huygens source can be obtained through a simple expression given in the next section. The radiation characteristics of the antenna can be found by adding or integrating the effects due to all Huygens sources of the wave front.

This procedure is based on the assumption that the antenna is transmitting. A basic law of reciprocity assures us that the receiving gain and radiation characteristics of the antenna will be identical with the transmitting ones when only linear elements are involved.

This resolution of an antenna wave front into an array of secondary sources can be justified within certain limitations on the basis of the induction theorem of electromagnetic theory.[3] These limitations are discussed in a qualitative way in section 3.13.

3.1 *The Huygens Source*

Consider an elementary Huygens source of electric intensity E_0 polarized parallel to the X axis with area dS in the XY plane (Fig. 8). This can be thought of as an element of area dS of a wave front of a linearly polarized plane electromagnetic wave which is advancing in the positive z direction.[3] From Maxwell's Equations we can determine the field at any point of space due to this Huygens Source. The components of electric field, are found to be

$$\begin{aligned} E_r &= 0 \\ E_\theta &= i \frac{E_0\, dS}{2\lambda r}\, e^{-i(2\pi/\lambda)r}\,(1 + \cos\theta)\cos\phi \\ E_\phi &= -i \frac{E_0\, dS}{2\lambda r}\, e^{-i(2\pi/\lambda)r}\,(1 + \cos\theta)\sin\phi \end{aligned} \tag{15}$$

where λ is the wavelength.

We see at once that this represents a vector whose absolute magnitude at all points of space is given by

$$|E| = \frac{E_0\, dS}{2\lambda r}\,(1 + \cos\theta). \tag{16}$$

[3] S. A. Schelkunoff, Loc. Cit., Chap. 9.

Here $\frac{E_0 dS}{\lambda}$ is an amplitude factor which depends on the wavelength, intensity and area of the elementary source and $1/r$ is an amplitude factor which specifies the variation of field with distance. $(1 + \cos \theta)$ is an amplitude factor which shows that the directional pattern of the elementary source is a cardioid with maximum radiation in the direction of propagation and no radiation in the reverse direction.

When we use the properties of the Huygens source in analyzing a micro-

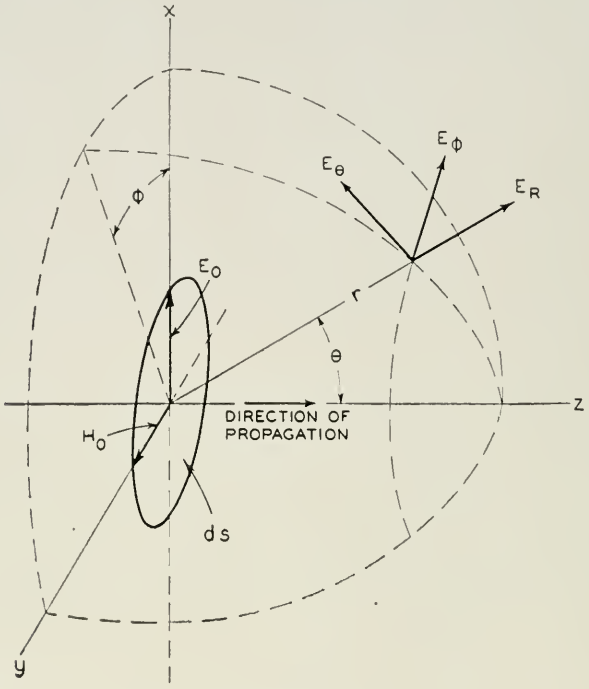


Fig. 8—The Huygens Source.

wave antenna we are usually concerned principally with radiation in or near the direction of propagation. For such radiation Equation 16 takes a particularly simple form in Cartesian Coordinates

$$E_x \cong i \frac{E_0 dS}{\lambda r} e^{-i(2\pi/\lambda)r}; \; E_y \cong 0; \; E_z \cong 0. \tag{17}$$

This represents an electric vector nearly parallel to the electric vector of the source. The amplitude is given by the factor $\frac{E_0 dS}{\lambda r}$ and the phase by the

factor $i\, e^{-i(2\pi/\lambda)r}$. With this equation as a basis we will now proceed to study some relevant matters concerning radar antennas.

3.2 *Gain and Effective Area of an Ideal Antenna*

On the basis of (17) we can now determine the gain of an ideal antenna of area S ($S \gg \lambda^2$). This antenna is assumed to be free of heat loss and to transmit by generating an advancing wave which is uniform in phase and amplitude in the XY plane. Let the electric intensity in the wave front of

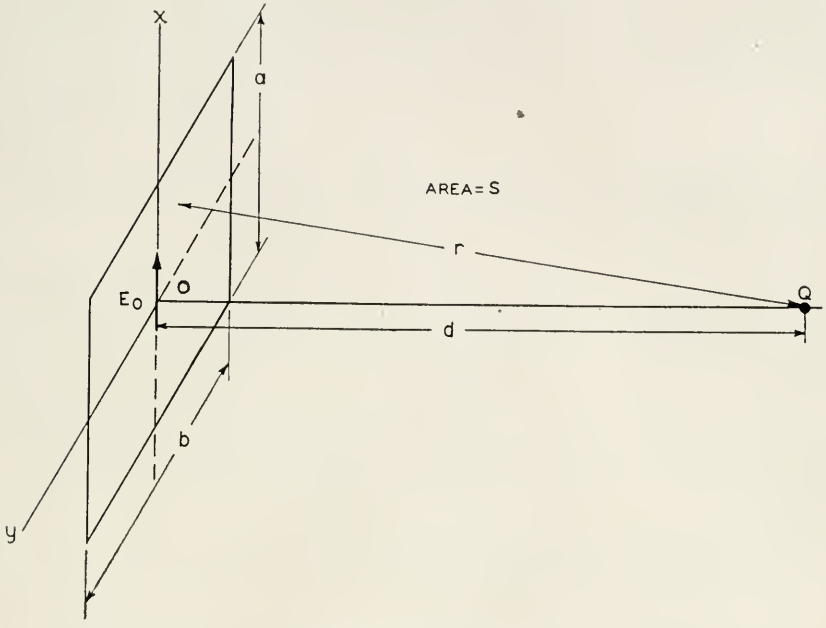


Fig. 9—An Ideal Antenna.

the ideal antenna be E_0 polarized parallel to the X axis (Fig. 9). The transmitted power P_T is equal to the power flow through S and is given by

$$P_T = \frac{E_0^2}{120\pi} S. \tag{18}$$

At a point Q on the Z axis the electric intensity is obtained by adding the effects of all the Huygens sources in S. If the distance of Q from O is so great that

$$r = d + \Delta$$

where Δ is a negligibly small fraction of a wavelength for every point on S then we see from (17) that the electric vector at Q is given by

$$E_x = \int_S i \frac{E_0 \, dS}{\lambda r} e^{-i(2\pi/\lambda)r} = i e^{-i(2\pi/\lambda)d} \frac{E_0 S}{\lambda d}; \; E_y = 0; \; E_z = 0. \tag{19}$$

The power flow per unit area at Q is therefore

$$P = \frac{1}{120\pi} \frac{E_0^2 S^2}{\lambda^2 d^2} = \frac{P_T S}{\lambda^2 d^2}$$

P_0 the power flow per unit area at Q when power is radiated isotropically from O is found by assuming that P_T is spread evenly over the surface of a sphere of radius d.

$$P_0 = \frac{P_T}{4\pi d^2} \tag{20}$$

The gain of a lossless, uniphase, uniamplitude, linearly polarized antenna is, by the definition of equation 1, the ratio of 19 and 20.

$$G_s = \frac{P_T S}{\lambda^2 d^2} \Big/ \frac{P_T}{4\pi d^2} = \frac{4\pi S}{\lambda^2}. \tag{21}$$

It follows from 12 that the effective area of the ideal antenna is

$$A = S \tag{22}$$

In other words in this ideal antenna the effective area is equal to the actual area. This is a result which might have been obtained by more direct arguments.

3.3 *Gain and Effective Area of an Antenna with Aperture in a Plane and with Arbitrary Phase and Amplitude*

Let us consider an antenna with a wave front in the XY plane which has a known phase and amplitude variation. Let the electric intensity in the wave front be

$$E(x, y) = E_0 a(x, y) e^{i\phi(x,y)} \tag{23}$$

polarized parallel to the x axis. The radiated power is equal to the power flow through S and is given by

$$P_{\text{rad}} = \frac{E_0^2 \int a^2(x, y) \, dS}{120\pi}. \tag{24}$$

The input power to the antenna is

$$P_T = P_{\text{rad}}/L \tag{25}$$

where L is a loss factor (<1). At a point Q on the Z axis the electric intensity is obtained by adding the effects of all the Huygens sources in S. If OQ is as great as in the above derivation for the gain of an ideal antenna then we see from 17 that the electric intensity at Q is

$$E_x = i \frac{e^{i(2\pi/\lambda)d}}{\lambda d} E_0 \int a(x, y) e^{i\phi(x,y)} \, dS; \; E_y = 0; \; E_z = 0. \tag{26}$$

The power flow per unit area at Q is given by

$$P = \frac{1}{120\pi} |E_x|^2 \tag{27}$$

and P_0 the power flow per unit area at Q when P_T is radiated isotropically is given by equation (3).

The power gain of the antenna, by definition 1 is therefore

$$G = \frac{P}{P_0} = \frac{|E_x|^2}{120\pi} \Big/ \frac{P_T}{4\pi d^2} = \frac{4\pi L}{\lambda^2} \frac{\left| \int_S a(x, y) e^{i\phi(x,y)} \, dS \right|^2}{\int_S a^2(x, y) \, dS}. \tag{28}$$

The gain expressed in *db* is given by

$$G_{db} = 10 \log_{10} G \tag{29}$$

We combine 12 and 28 to obtain

$$A = L \frac{\left| \int_S a(x, y) e^{i\phi(x,y)} \, dS \right|^2}{\int_S a^2(x, y) \, dS} \tag{30}$$

a formula for the effective area of the antenna.

3.4 *The Significance of the Pattern of a Radar Antenna*

The accuracy with which a radar can determine the directions to a target depends upon the *beam widths* of the radar antenna. The ability of the radar to separate a target from its background or distinguish it from other targets depends upon the beam widths and the *minor lobes* of the radar antenna. The efficiency with which the radar uses the available power to view a given region of space depends on the *beam shape* of the antenna. These quantities characterize the antenna pattern. In the following sections means for the calculation of antenna patterns in terms of wave front theory will be developed, and some illustrations will be given.

3.5 *Pattern in Terms of Antenna Wave Front*

If the relative phase and amplitude in a wave front are given by

$$E(x, y) = a(x, y)e^{i\phi(x,y)} \tag{31}$$

the relative phase and amplitude at a distant point Q not necessarily on the Z axis (Fig. 10) in the important case where the angle QOZ between the direction of propagation and the direction to the point is small, is given from (17) by adding the contributions at Q due to all parts of the wave front. This gives

$$E_Q = \frac{i}{\lambda d}\int_s e^{-i(2\pi/\lambda)r}\, e^{i\phi(x,y)}\, a(x, y)\, dS. \tag{32}$$

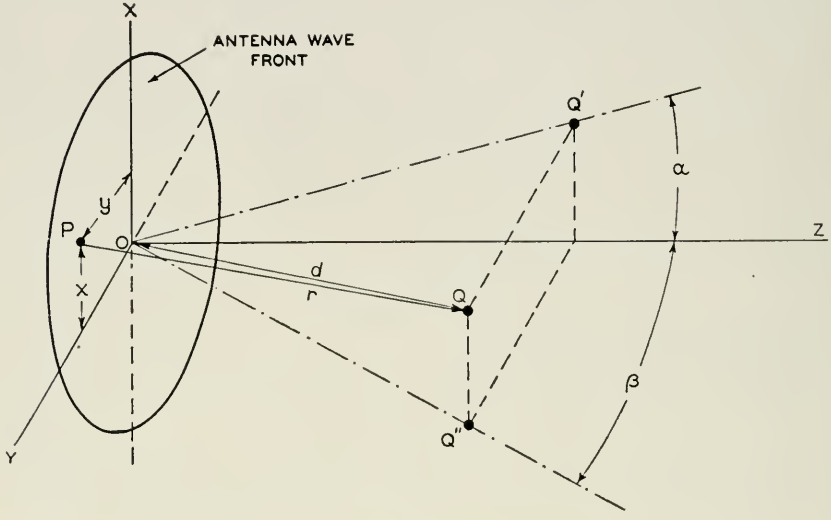


Fig. 10—Geometry of Pattern Analysis.

The quantity r in (32) is the distance from any point P with coordinates x, y, O, in the XY, plane to the point Q (Fig. 10). Simple trigonometry shows that when OQ is very large

$$r = d - x\sin\alpha - y\sin\beta \tag{33}$$

where d is the distance OQ, α is the angle ZOQ' between OZ and OQ' the projection of OQ on the XZ plane and β is similarly the angle ZOQ''. The substitution of 33 into 32 gives

$$E_Q = \frac{i\, e^{-i(2\pi/\lambda)d}}{\lambda d}\int_s e^{i(2\pi/\lambda)(x\sin\alpha + y\sin\beta) + i\phi(x,y)}\, a(x, y)\, dS. \tag{34}$$

In most practical cases this equation can be simplified by the assumptions

$$\phi(x, y) = \phi'(x) + \phi''(y)$$
$$a(x, y) = a'(x)a''(y)$$

from which it follows that

$$| E_Q | = F(d)F(\alpha)F(\beta) \tag{35}$$

where $F(d)$ is an amplitude factor which does not depend on angle,

$$F(\alpha) = \int e^{i(2\pi/\lambda)x\sin\alpha + i\phi'(x)} a'(x)\, dx \tag{36}$$

is a directional factor which depends only on the angle α and not on the angle β or d, and $F(\beta)$ similarly depends on β but not on α or d. The pattern of an antenna can be calculated with the help of the simple integrals as in 36, and illustrations of such calculations will be given in the following sections.

3.6 *Pattern of an Ideal Rectangular Antenna*

Let the wave front be that of an ideal rectangular antenna of dimensions a, b; with linear polarization and uniform phase and amplitude. The dimensions a and b can be placed parallel to the X and Y axes respectively as sketched in Fig. 9. Equation 36 then gives

$$F(\alpha) = \int_{-a/2}^{a/2} e^{i(2\pi/\lambda)x\sin\alpha}\, dx = a\,\frac{\sin\psi}{\psi} \tag{37}$$

where $\psi = \dfrac{\pi a \sin\alpha}{\lambda}$.

Similarly

$$F(\beta) = b\,\frac{\sin\psi'}{\psi'} \tag{38}$$

where $\psi' = \dfrac{\pi b \sin\beta}{\lambda}$.

The pattern of the ideal rectangular aperture, in other words the distribution of electrical field in angle is thus given approximately by

$$F(\alpha)F(\beta) = ab\,\frac{\sin\psi}{\psi}\,\frac{\sin\psi'}{\psi'}. \tag{39}$$

The function $\dfrac{\sin\psi}{\psi}$ is plotted in Fig. 11. It is perhaps the most useful function of antenna theory, not because ideal antennas as defined above are particularly desirable in practice but because they provide a simple stand-

ard with which more useful but more complex antennas can profitably be compared.

3.7 *Effect on Pattern of Amplitude Taper*

The $\frac{\sin \psi}{\psi}$ pattern which results from an ideal wave front has undesirably high minor lobes for most radar applications. These minor lobes will be reduced if the wave front of constant amplitude is replaced by one which retains a constant phase but has a rounded or 'tapered' amplitude distribution.

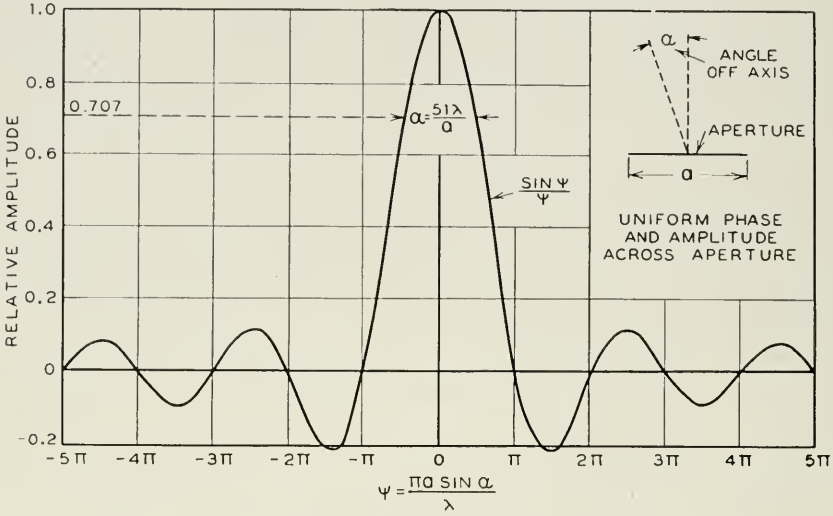


Fig. 11—Pattern of Ideal Rectangular Antenna.

If such an amplitude taper is represented analytically by the function

$$a'(x) = C_1 + C_2 \cos \frac{\pi x}{a} \tag{40}$$

then equation (36) is readily integrable. To integrate it we utilize the identity

$$\cos \frac{\pi x}{a} = \frac{e^{+i\pi x/a} + e^{-i\pi x/a}}{2}$$

upon which the integral becomes the sum of three simple integrals of the form

$$\int_{-a/2}^{a/2} e^{ikx} dx = a \left[\frac{\sin \frac{ka}{2}}{\frac{ka}{2}} \right] \tag{41}$$

We therefore obtain

$$F(\alpha) = aC_1 \frac{\sin \psi}{\psi} + a \frac{C_2}{2} \left[\frac{\sin\left(\psi + \frac{\pi}{2}\right)}{\left(\psi + \frac{\pi}{2}\right)} + \frac{\sin\left(\psi - \frac{\pi}{2}\right)}{\left(\psi - \frac{\pi}{2}\right)} \right] \tag{42}$$

The patterns resulting from two possible tapers are given by substituting $C_1 = 0, C_2 = 1$ and $C_1 = 1/3, C_2 = 2/3$ in (42). These patterns are evidently calculable in terms of the known function $\frac{\sin \alpha}{\alpha}$. They are plotted in Figs. 12 and 13.

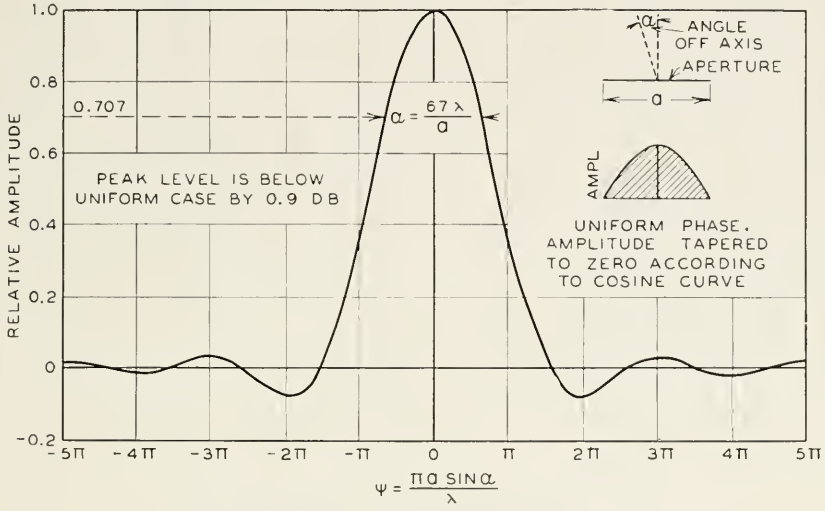


Fig. 12—Pattern of Tapered Rectangular Antenna.

It will be observed that minor lobe suppression through tapering is obtained at the expense of beam broadening. In addition to this the gain is reduced by tapering, as could have been calculated from 28. These undesirable effects must be contended with in any practical antenna design. The choice of taper must be made on the basis of the most desirable compromise between the conflicting factors.

3.8 *Effect on Pattern of Linear Phase Variation*

If we assume a constant amplitude and a linear phase variation

$$\phi'(x) = -k_1 x$$

over an aperture $-a/2 < x < a/2$ then 36 becomes a simple integral of the form (41) and we obtain

$$F(\alpha) = a \frac{\sin \psi''}{\psi''} \quad \text{where} \quad \psi'' = \frac{\pi a}{\lambda} \sin \alpha - \frac{k_1 a}{2} \tag{43}$$

The physical interpretation of (43) is simply that the pattern is identical to the pattern of an antenna with constant amplitude and uniform phase but rotated through an angle θ where

$$\sin \theta = \frac{k_1 \lambda}{2\pi}$$

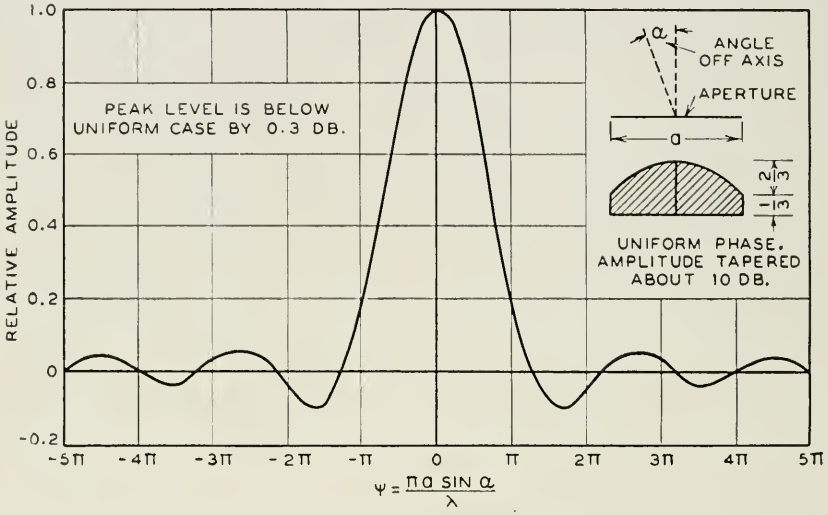


Fig. 13—Pattern of Tapered Rectangular Antenna.

Simple examination shows that the new direction of the radiation maximum is at right angles to a uniphase surface, as we would intuitively expect. This phenomenon has particular relevance to the design of scanning antennas.

3.9 *Effect on Pattern of Square Law Phase Variation*

If we assume a constant amplitude and a square law phase variation

$$\phi'(x) = -k_2^2 x^2$$

over the aperture $a/2 < x < a/2$ then the substitution

$$x = \frac{1}{k_2}\left[X + \frac{\frac{2\pi}{\lambda} \sin \alpha}{2k_2}\right] \tag{44}$$

reduces (36) to the form

$$F(\alpha) = \frac{1}{k_2} e^{i\left(\frac{2\pi}{\lambda}\right)^2 \sin^2 \alpha / 4k_2^2} \int e^{-iX^2} dX \tag{45}$$

Equation (45) can be evaluated with the help of Fresnel's Integrals

$$\int \cos X^2 \, dX, \qquad \int \sin X^2 \, dX$$

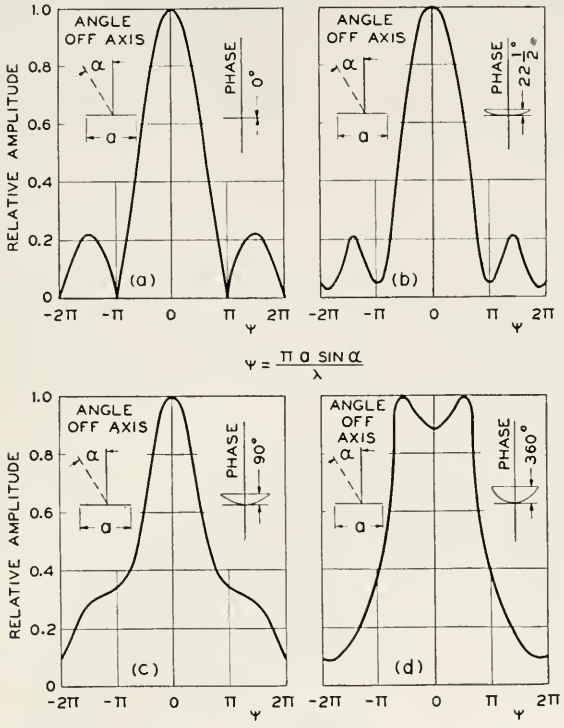


Fig. 14—Patterns of Rectangular Apertures with Square Law Phase Variation.

which are tabulated[4], or from Cornu's Spiral which is a convenient graphical representation of the Fresnel Integrals.

Typical computed patterns for apertures with square law phase variations are plotted in Fig. 14. These theoretical curves can be applied to the following important practical problems.

(1) The pattern of an electromagnetic horn.

[4] For numerical values of Fresnel's Integrals and a plot of Cornu's Spiral see Jahnke and Emde, *Tables of Functions* B. G. Teubner, Leipzig, 1933, or Dover Publications, New York City, 1943.

(2) The defocussing of a reflector or lens due to improper placing of the primary feed.
(3) The defocussing of a zoned reflector or lens due to operation at a frequency off mid-band.

In addition to providing distant patterns of apertures with curved wave fronts (44) provides theoretical 'close in' patterns of antennas with plane wave fronts. This arises from the simple fact that a plane aperture appears as a curved aperture to close in points. The degree of curvature depends on the distance and can be evaluated by extremely simple geometrical considerations. When this has been done we find that Fig. 14 represents the so-called Fresnel diffraction field.

With this interpretation of square law variation of the aperture we can examine several additional useful problems. We can for instance justify the commonly used relation

$$D = \frac{2a^2}{\lambda}$$

for the minimum permissible distance of the field source from an experimental antenna test site. This distance produces an effective phase curvature of $\lambda/16$. We can examine optical antenna systems employing large primary feeds, in particular those employing parabolic cylinders illuminated by line sources.

3.10 *Effect on Pattern of Cubic Phase Variation*

If we assume a constant amplitude and a cubic phase variation $\phi'(x) = -k_3^3 x^3$ over the aperture from $-a/2 < x < a/2$ then equation (36) becomes

$$F(\alpha) = \int_{-a/2}^{a/2} e^{-ik_3^3 x^3} \cdot e^{i(2\pi/\lambda)x \sin\alpha} \cdot dx \tag{46}$$

If $k_3^3 x^3 \leq \frac{\pi}{2}$ then it is a fairly good approximation to write

$$e^{-ik_3^3 x^3} = 1 - ik_3^3 x^3 - \frac{k_3^6 x^6}{2} + \cdots \tag{47}$$

from which it follows that (46) can be integrated since it reduces to a sum of three terms each of which can be integrated.

Typical computed patterns for apertures with cubic phase variation are plotted in Figs. 15 and 16. Cubic phase distortions are found in practice when reflectors or lenses are illuminated by primary feeds which are off axis either because of inaccurate alignment or because beam lobing or scanning through feed motion is desired. The beam distortion due to cubic phase variation is known in optics as 'coma' and the increased unsymmetrical lobe which is particularly evident in Fig. 16 is commonly called a 'coma lobe'.

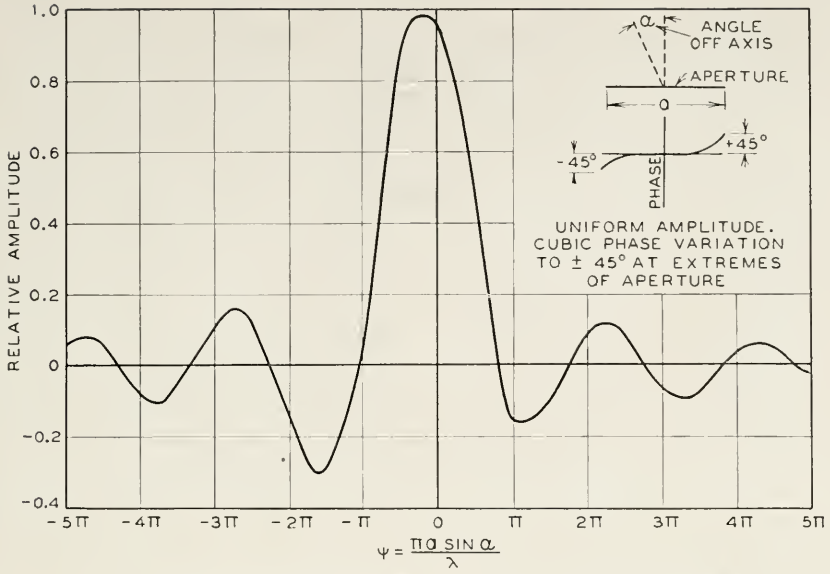


Fig. 15—Pattern of Rectangular Antenna with Cubic Phase Variation.

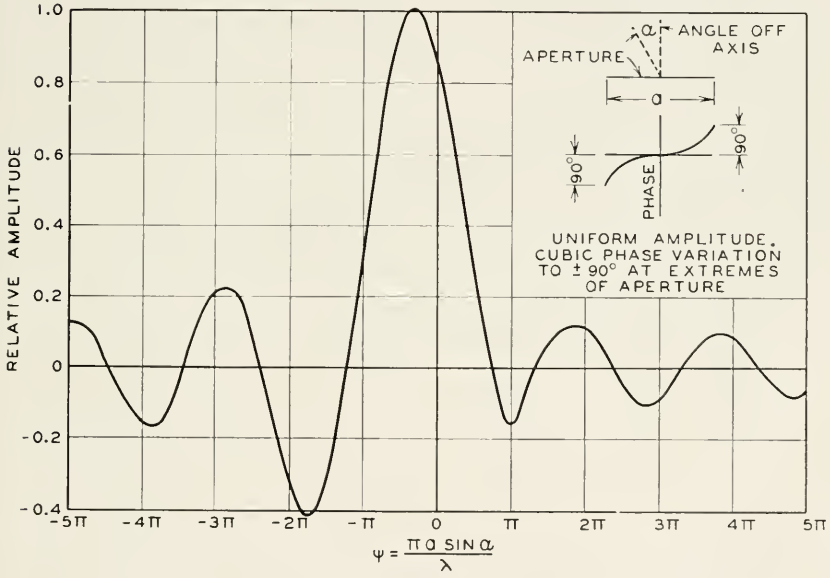


Fig. 16—Pattern of Rectangular Antenna with Cubic Phase Variation.

3.11 *Two General Methods*

In sections 3.7 and 3.8 we integrated (36) by expressing $a'(x)e^{i\phi'(x)}$ as a sum of terms of the form e^{ikx}. Since $a'(x)e^{i\phi'(x)}$ for finite amplitudes in a finite

aperture can always be expressed as a Fourier sum of this form this solution can in principle always be found.

Alternatively in section 3.10 the integral was evaluated as a sum of integrals of the general type $\int x^n e^{ikx}\,dx$. Since $a'(x)e^{i\phi'(x)}$ for finite amplitudes in a finite aperture can always be expressed in terms of a power series, this solution can also in principle always be found.

3.12 *Arrays*

When the aperture consists of an array of component or unit apertures the evaluation of (36) must be made in part through a summation. When all of the elementary apertures are alike this summation can be reduced to the determination of an 'Array Factor'. The pattern of the array is given by multiplying the array factor by the pattern of a single unit.

The pattern of an array of identical units spaced equally at distances somewhat less than a wavelength can be proved to be usually almost equivalent to the pattern of a continuous wave front with the same average energy density and phase in each region.

3.13 *Limitations to Antenna Wave Front Analysis*

Through the analysis of antenna characteristics by means of wave front theory as based on equation (17) we have been able to demonstrate some of the fundamental theoretical principles of antenna design. The use of this simple approach is justified fully by its relative simplicity and by its applicability to the majority of radar antennas. Nevertheless it cannot always be used. It will certainly be inaccurate or inapplicable in the following cases:

(1) When any dimension of the aperture is of the order of a wavelength or smaller (as in many primary feeds).
(2) Where large variations in the amplitude or phase in the aperture occur in distances which are of the order of a wavelength or smaller (as in dipole arrays).
(3) Where the antenna to be considered does not act essentially through the generation of a plane wave front (as in an end fire antenna or a cosecant antenna).

When the wave front analysis breaks down alternative satisfactory approaches based on Maxwell's equation are sometimes but not always fruitful. Literature on more classical antenna theory is available in a variety of sources. For much fundamental and relevant theoretical work the reader is referred to Schelkunoff.[5]

[5] S. A. Schelkunoff, Loc. Cit.

4. Application of General Principles

In the foregoing sections we have provided some discussion of what happens to a radar signal from the time that the pulse enters the antenna on transmission until the time that the reflected signal leaves the antenna on reception. We have for convenience divided the principles which chiefly concern us into three groups, transmission line theory, transmission principles and wave front theory.

With the aid of transmission line theory we can examine problems concerning locally guided or controlled energy. The details of the problems of antenna construction, such as those to be discussed in Part II frequently demand a grasp of transmission line theory. With it we can study local losses, due to resistance or leakage, which affect the gain of the antenna. We can examine reflection problems and their effect on the match of the antenna. Special antennas, such as those employing phase shifters or transmission between parallel conducting plates, introduce many special problems which lie wholly or partly in the transmission line field.

An understanding of the principles which govern transmission through free space aids us in comprehending the radar antenna field as a whole. Through a general understanding of antenna gains and effective areas we are better equipped to judge their significance in particular cases, and to evaluate and control the effects of particular methods of construction on them.

Wave front theory provides us with a powerful method of analysis through which we can connect the radiation characteristics produced by a given antenna with the radiating currents in the antenna. Through it we can examine theoretical questions concerning beam widths and shape, unwanted radiation and gain.

An understanding of theory is necessary to the radar antenna designer, but it is by no means sufficient. It is easy to attach too much importance to theoretical examination and speculation while neglecting physical facts which can 'make or break' an antenna design. Theory alone provides no substitute for the practical 'know how' of antenna construction. It cannot do away with the necessity for careful experiment and measurement. Least of all can it replace the inventiveness and aggressive originality through which new problems are solved and new techniques are developed.

PART II

METHODS OF ANTENNA CONSTRUCTION

5. General

Techniques are essential to technical accomplishment. An understanding of general principles alone is not enough. The designing engineer must have

at his disposal or develop practical methods which can produce the results he requires. The effectiveness and simplicity of these methods are fair measures of the degree of technical development.

The study of methods of radar antenna construction is the study of the means by which radar antenna requirements are met. In a broader sense this includes an examination of mechanical structures, of the metals and plastics from which antennas are made, of the processes by which they are assembled, and of the finishes by which they are protected from their environment. It might include a study of practical installation and maintenance procedures. But these matters, which like the rest of Radar have unfolded widely during the war, are beyond the scope of this paper. An adequate discussion of them would have to be based on hundreds of technical reports and instruction manuals and on thousands of manufacturing drawings. The account of methods which is to follow will therefore be restricted to a discussion, usually from the electrical point of view, of the more useful and common radar antenna configurations.

6. Classification of Methods

During the history of radar, short as it is, many methods of antenna construction have been devised. To understand the details of all of these methods and the diverse applications of each is a task that lies beyond the ability of any single individual. Nevertheless most of the methods fall into one or another of a limited collection of groups or classifications. We can grasp most of what is generally important through a study of these groups.

In order to provide a basis for classification we will review briefly, from a transmitting standpoint, the action of an antenna. Any antenna is in a sense a transformer between a transmission line and free space. More explicitly, it is a device which accepts energy incident at its terminals, and converts it into an advancing electromagnetic wave with prescribed amplitude, phase and polarization over an area. In order to do this the antenna must have some kind of energy distributing system, some means of amplitude control and some means of phase control. The distributed energy must be suitably controlled in phase, amplitude and polarization.

All antennas perform these functions, but different antennas perform them by different means. Through an examination of the means by which they are performed and the differences between them we are enabled to classify methods of antenna construction.

To distribute energy over its aperture an antenna can use a branching system of transmission lines. When this is done the antenna is an array. Arrays are particularly common in the short wave communication bands, but somewhat less common in the microwave radar bands. In a somewhat simpler method the antenna distributes energy over an area by radiating it

from an initial source or 'primary feed'. This distribution can occur in both dimensions at once, as from a point source. Alternatively the energy can be radiated from a primary source but be constrained to lie between parallel conducting plates so that it is at first distributed only over a long narrow aperture or 'line source'. Distribution over the other dimension occurs only after radiation from the line source.

In order to control the amplitude across the aperture of an array antenna we must design the branching junctions so that the desired power division occurs in each one. When the energy is distributed by radiation from a primary source we must control the amplitude by selecting the proper primary feed directivity.

We can control the phase of an array antenna by choosing properly the lengths of the branching lines. Alternatively we can insert appropriate phase changers in the lines.

When the energy is distributed by primary feeds, methods resembling those of optics can be used to control phase. The radiation from a point source is spherical in character. It can be 'focussed' into a plane wave by means of a paraboloidal reflector or by a spherical lens. The radiation from a point source between parallel plates or from a uniphase line source is cylindrical in character. It can be focussed by a parabolic cylinder or a cylindrical lens.

In Table A we have indicated a possible classification of methods of radar antenna construction. This classification is based on the differences discussed in the foregoing paragraphs.

TABLE A

CLASSIFICATION OF METHODS OF RADAR ANTENNA CONSTRUCTION

- Methods of Radar Antenna Construction
  - Arrays of
    - Dipoles
    - Polyrods
    - Optical Elements
  - Optical Methods
    - Spherical Optics
      - Point sources and Spherical Elements
        - Dipole Arrays
        - Wave Guide Apertures
        - Reflectors
        - Lenses
    - Cylindrical Optics
      - Line sources and Cylindrical Elements
        - Arrays
        - Reflectors
        - Lenses
        - Reflectors
        - Lenses

7. Basic Design Formulation

Certain design factors are common to almost all radar antennas. Because of their importance it would be well to consider these factors in a general way before proceeding with a study of particular antenna techniques.

Almost every radar antenna, regardless of how it is made, has a well defined aperture or wave front. Through wave front analysis we can often examine the connections between the Huygens sources in the antenna aperture and the radiation characteristics of the antenna. We can, in other words, use wave front analysis to study the fundamental antenna design factors, provided the analysis does not violate one of the conditions of section 3.13.

7.1 *Dimensions of the Aperture*

The dimensions of the aperture of a properly designed antenna are related to its gain by simple and general approximate relations. If the aperture is uniphase and has an amplitude distribution that is not too far from constant the relation

$$G = \frac{4\pi A}{\lambda^2}$$

is useful in connecting the gain of an antenna with the area of its aperture. The effective area is related to the area of the aperture by the equation

$$A = \eta S$$

where η is an efficiency factor. In principle η could have any value but in practice for microwave antennas η has always been less than one. Its value for most uniphase and tapered amplitude antennas is between 0.4 and 0.7. In special cases, e.g. for cosecant antennas or for some scanners its value may be less than 0.4.

The necessary dimensions for the aperture may be determined from the required beam widths in two perpendicular directions. Beam widths are usually specified as half power widths, that is by the number of degrees between directions for which the one way response is 3 *db* below the maximum response. Figure 11 shows that for an ideal rectangular antenna with uniform phase, polarization and amplitude $\alpha_{P/2} = 51 \frac{\lambda}{a}$ degrees where $\alpha_{P/2} =$ half power width in degrees, $a =$ aperture dimension and $\lambda =$ wavelength. The relation $\alpha_{P/2} = 65 \frac{\lambda}{a}$ degrees is more nearly correct for the majority of practical antennas with round or elliptical apertures and with uniform phase and reasonably tapered amplitudes.

7.2 *Amplitude Distribution*

Except where special, in particular cosecant, patterns are desired the principle factors affecting amplitude distribution are efficiency and required minor lobe level. The amplitude distribution or taper of an ideal uniphase rectangular wave front affects the minor lobe level as indicated by Figures 11, 12 and 13. Practical antennas tend to fall somewhat below this ideal picture because of non-uniform phase and because of variations from the ideal amplitude distribution due to discontinuities in the aperture and undesired leakage or spillover of energy. Nevertheless a commonly used rule of thumb is that minor lobes 20 *db* or more below the peak radiation level are tolerable and will not be exceeded with a rounded amplitude taper of 10 or 12 *db*.

7.3 *Phase Control*

Uniphase wave fronts are used whenever a simple pattern with prescribed gain, beam widths and minor lobes is to be obtained with minimum aperture dimensions. When special results are desired such as cosecant patterns or scanning beams the phase must be varied in special ways.

Mechanical tolerances in the antenna structure make it impossible to hold phases precisely to the desired values. The accuracy with which the phases can be held constant in practice varies with the technique, the antenna size and the wave length. Undesired phase variations increase the minor lobes and reduce the gain of an antenna. The extent to which phase variations can be expected to reduce the gain is indicated in Fig. 17.

8. Parabolic Antennas

The headlights of a car or the searchlights of an antiaircraft battery use reflectors to produce beams of light. Similarly the majority of radar antennas employ reflectors to focus beams of microwave energy. These reflectors may be exactly or approximately parabolic or they may have special shapes to produce special patterns. If they are parabolic they may be paraboloids which are illuminated by point sources and focus in both directions, or they may be parabolic cylinders which focus in only one direction. If they are parabolic cylinders they may be illuminated by line sources or they may be confined between parallel conducting plates and illuminated by point sources to produce line sources.

8.1 *Control of Phase*

A simple and natural way to distribute energy smoothly in space is to radiate it from a relatively nondirectional 'primary' source such as a dipole array or an open ended wave guide. This energy will be formed into a directive beam if a reflector is introduced to bring it to a plane area or wave front with constant phase. If the primary source is effectively a point as far as

phase is concerned, that is if the radiated energy has the same phase for all points which are the same distance from a given point, then the reflector should be parabolic. This can be proved by simple geometrical means.

In Fig. 18 let the point source S coincide with the point $x = f$, $y = O$ of a coordinate system and let the uniphase wave front coincide with the line $x = f$. Let us assume that one point O of the reflector is at the origin. Then it can be shown that any other point of the reflector must lie on the curve

$$y^2 = 4fx$$

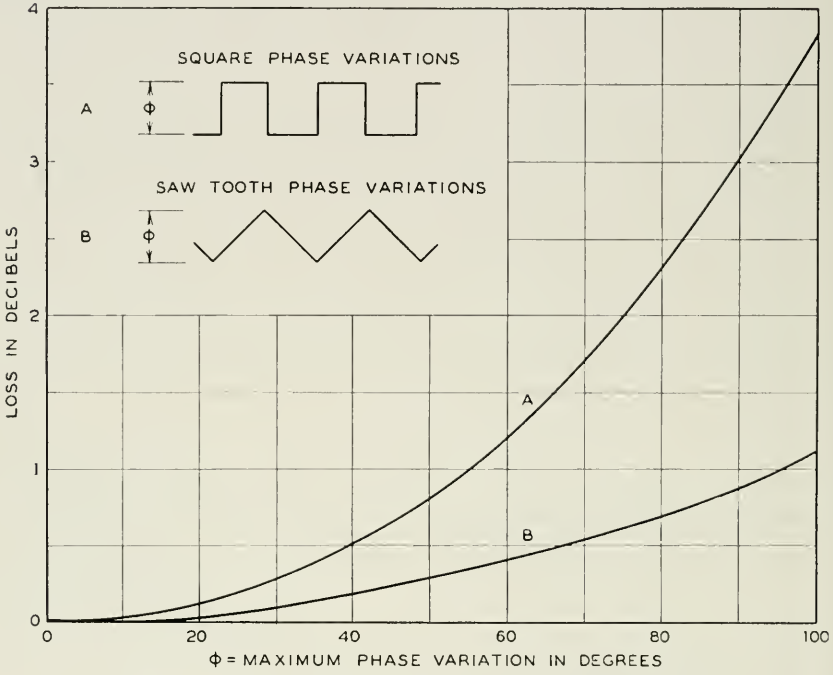


Fig. 17—Loss due to Phase Variation in Antenna Wave Front.

This is a parabola with focus at f, o and focal length f.

The derivation based on Fig. 18 is two dimensional and therefore in principle applies as it stands only to line source antennas employing parabolic cylinders bounded by parallel conducting planes (Fig. 24 and 25). If Fig. 18 is rotated about the X axis the parabola generates a paraboloid of revolution (Fig. 3). This paraboloid focusses energy spreading spherically from the point source at S in such a way that a uniphase wave front over a plane area is produced. Alternatively Fig. 18 can be translated in the Z direction perpendicular to the XY plane. The parabola then generates a

parabolic cylinder and the point source S generates a line source at the focal line of the parabolic cylinder (Fig. 19). The energy spreading cylindrically from the line source is focussed by the parabolic cylinder in such a way that a uniphase wave front over a plane area is again produced. Parabolic cylinders and paraboloids are both used commonly in radar antenna practice.

In the discussion so far it has been assumed that the primary source is effectively a point source and that the reflector is exactly parabolic. If the primary source is not effectively a point source, in other words if it produces waves which are not purely spherical, then the reflector must be distorted from the parabolic shape if it is to produce perfect phase correction. When

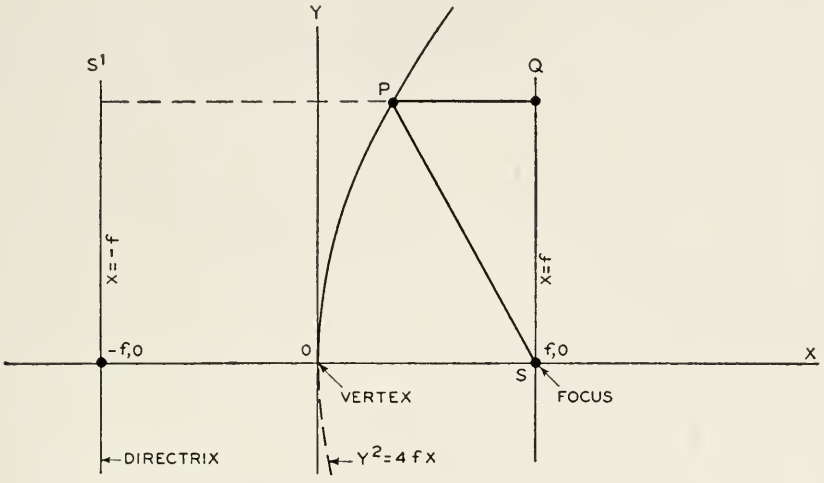


Fig. 18—Parabola.

this occurs the correct reflector shape is sometimes specified on the basis of an experimental determination of phase.

8.2 *Control of Amplitude*

When a primary source is used to illuminate a parabolic reflector there are two factors which affect the amplitude of the resulting wave front. One of these is of course the amplitude pattern of the primary source. The other is the geometrical or space attenuation factor which is different for different parts of the wave front. In most practical antennas each of these factors tends to taper the amplitude so that it is less at the edges of the antenna than it is in the central region. The effective area of the antenna is reduced by this taper.

In any finite parabolic antenna some of the energy radiated by the primary

source will fail to strike the reflector. The effective area must also be reduced by the loss of this 'spill-over' energy.

The maximum effective area for a parabolic antenna is obtained by designing the primary feed to obtain the best compromise between loss due to taper and loss due to spill-over. It has been shown theoretically that this best compromise generally occurs when the amplitude taper across the aperture is about 10 or 12 *db* and that in the neighborhood of the optimum the efficiency is not too critically dependent on the taper.[6]

This theoretical result is well justified by experience and has been applied to the majority of practical parabolic antennas. It applies both when the reflector is paraboloidal so that taper in both directions must be considered

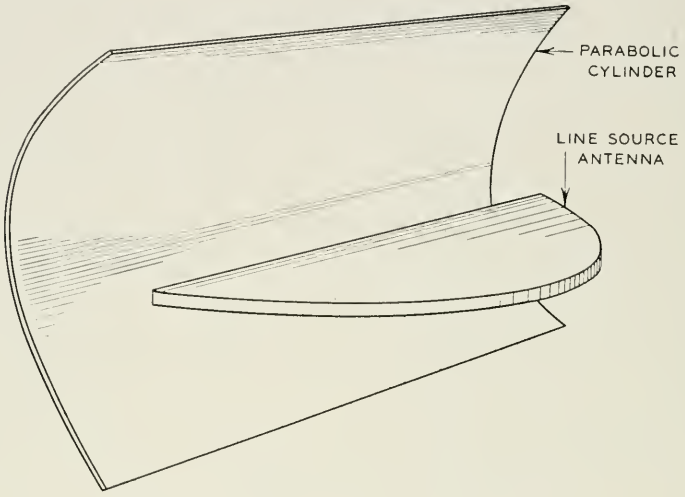


Fig. 19—A Parabolic Cylinder with Line Feed.

and when the reflector is a parabolic cylinder with only a single direction of taper. It is a fortunate by-product of a 10 or 12 *db* taper that it is generally sufficient to produce satisfactory minor lobe suppression.

8.3 *Choice of Configuration*

We have shown how a simple beam can be obtained through the use of a paraboloidal reflector with a point source or alternatively through the use of a reflecting parabolic cylinder and a line source. The line source itself can be produced with the help of a parabolic cylinder bounded by parallel conducting plates. We will now outline certain practical considerations. These considerations may determine which of the two reflector types will be

[6] C. C. Cutler, Parabolic Antenna Design for Microwaves, paper to be published in Proc. of the I. R. E.

used for a particular job. They may help in choosing a focal length and in determining which finite portion of a theoretically infinite parabolic curve should be used. Finally they may assist in determining whether reflector technique is really the best for the purpose at hand or whether we could do better with a lens or an array.

In designing a parabolic antenna it must obviously be decided at an early stage whether a paraboloid or one or more parabolic cylinders are to be employed. This choice must be based on a number of mechanical and electrical considerations. Paraboloids are more common in the radar art than parabolic cylinders and are probably to be preferred, yet a categorical a priori judgment is dangerous. It will perhaps be helpful to compare the two alternatives by the simple procedure of enumerating some features in which each is usually preferable to the other.

Paraboloidal antennas

(a) are simpler electrically, since point sources are simpler than line sources.
(b) are usually lighter.
(c) are more efficient.
(d) have better patterns in the desired polarization.
(e) are more appropriate for conical lobing or spiral scanning.

Antennas employing parabolic cylinders

(a) are simpler mechanically since only singly curved surfaces are required.
(b) have separate electrical control in two perpendicular directions.

This last advantage of parabolic cylinders is important in special antennas, many of which will be described in later sections. It is useful where antennas with very large aspect ratios (ratio of dimensions of the aperture in two perpendicular directions) are desired. It is highly desirable where control in one direction is to be achieved through some special means, as in cosecant antennas, or in antennas which scan in one direction only.

Let us suppose that we have selected the aperture dimensions and have decided whether the reflector is to be paraboloidal or cylindrical. The reflector is not yet completely determined for we are still free in principle to use any portion of a parabolic surface of any focal length. In order to obtain economy in physical size the focal length is generally made between 0.6a and 0.25a where 'a' is the aperture. For the same reason a section of the reflecting surface which is located symmetrically about the vertex is often chosen (Figures 3 and 19).

When a symmetrically located section of the reflector is used certain difficulties are introduced. These difficulties, if serious enough so that their removal justifies some increase in size can be bypassed through the use of an

offset section as shown in Fig. 20. We can comment on these difficulties as follows:

1. The presence of the feed in the path of the reflected energy causes a region of low intensity or 'shadow' in the wave front. The effect of this shadow on the antenna pattern depends on the size and shape of the feed and on the characteristics of the portion of the wave front where it is located. Its effect is to subtract from the undisturbed pattern a 'shadow pattern' component which is broad in angle. This decreases the gain and increases the minor lobes as indicated in Fig. 21.[7]

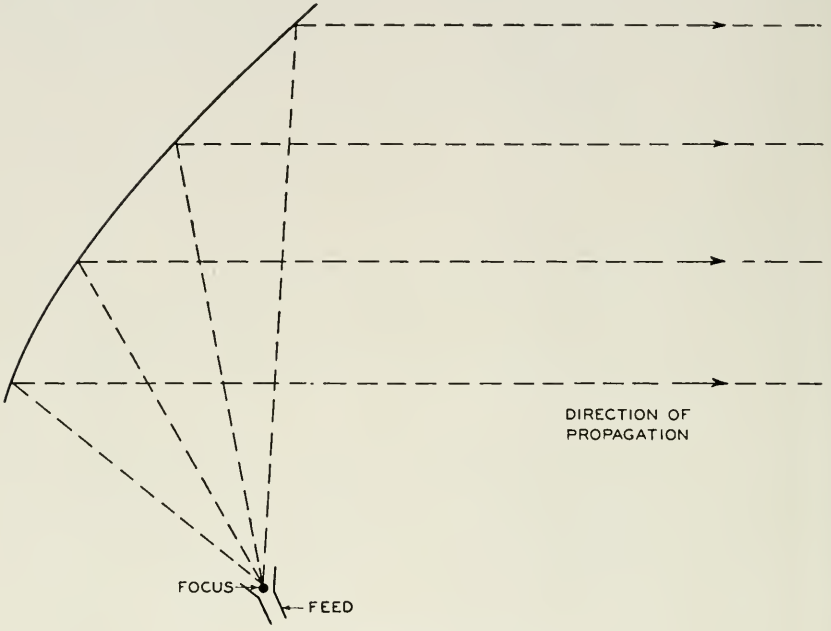


Fig. 20—Offset Parabolic Section.

2. Return of reflected energy into the feed introduces a standing wave of impedance mismatch in the feed line which is constant in amplitude but varies rapidly in phase as the frequency is varied. A mismatch at the feed which cancels the standing wave at one frequency will add to it at another frequency. A mismatch which will compensate over a band can be introduced by placing a raised plate of proper dimensions at the vertex of the reflector as indicated in Fig. 22, but such a plate produces a harmful effect on the pattern. In an antenna which must operate over a broad band it is consequently usually better to match

[7] Figures 21, 22, and 23 are taken from C. C. Cutler, loc. cit.

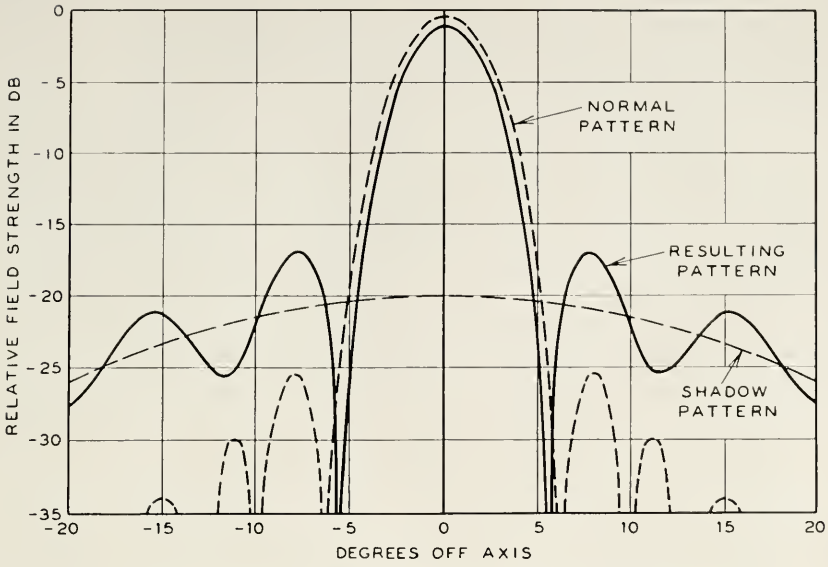


Fig. 21—Effect of Shadow on Paraboloid Radiation Pattern.

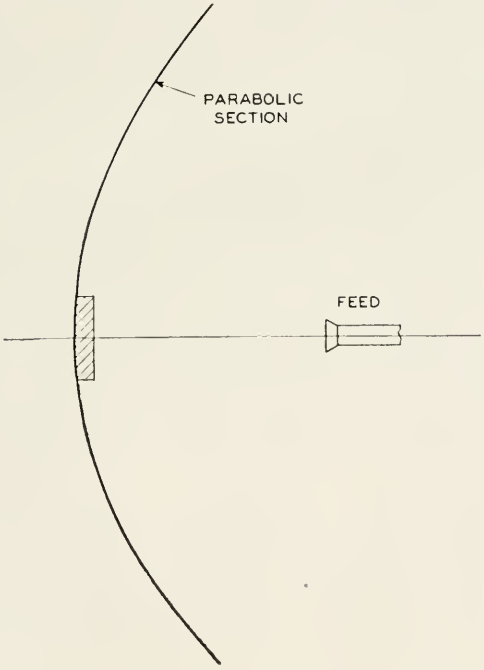


Fig. 22—Apex Matching Plate for Improving the Impedance Properties of a Parabola.

the feed to space and accept the residual standing wave, or if this is too great to use an offset section of the parabolic surface.

8.4 *Feeds for Paraboloids*

We have seen that an antenna with good wave front characteristics and consequently with a good beam and pattern can be constructed by illuminating a reflecting paraboloid with a properly designed feed placed at its focus. In this section we will examine the characteristics which the feed should have and some of the ways in which feeds are made in practice.

A feed for a paraboloid should

a. be appropriate to the transmission line with which it is fed. This is sometimes a coaxial line but more commonly a waveguide.
b. Provide an impedance match to this feed line. This match should usually be obtained in the absence of the reflector but sometimes, for narrow band antennas, with the reflector present.
c. have a satisfactory phase characteristic. For a paraboloid the feed should be, as far as phase is concerned, a true point source radiating spherical waves. As discussed at the end of 8.1, if the wave front is not accurately spherical, a compensating correction in the reflector can be made.
d. have a satisfactory amplitude characteristic. According to 8.2 this means that the feed should have a major radiation lobe with its maximum striking the center of the reflector, its intensity decreasing smoothly to a value about 8 to 10 *db* below the maximum in the direction of the reflector boundaries and remaining small for all directions which do not strike the reflector.
e. have a polarization characteristic which is such that the electric vectors in the reflected wave front will all be polarized in the same direction.
f. not disturb seriously the radiation characteristics of the antenna as a whole. The shadow effect of the feed, the feed line and the necessary mechanical supports must be small or absent. Primary radiation from the feed which does not strike the reflector or reflected energy which strikes the feed or associated structure and is then reradiated must be far enough down or so controlled that the antenna pattern is as required.

In addition to the electrical requirements for a paraboloid feed it must of course be so designed that all other engineering requirements are met, it must be firmly supported in the required position, must be connected to the antenna feed line in a satisfactory manner, must sometimes be furnished with an air tight or water tight seal, and so forth.

From the foregoing it is evident that a feed for a paraboloid is in itself a small relatively non-directive antenna. Its directivity is somewhat less than that obtained with an ordinary short wave array. It is therefore not surprising that dipole arrays are sometimes used in practice to feed paraboloids.

A simple dipole or half wave doublet can in itself be used to feed a paraboloid, but it is inefficient because of its inadequate directivity. It is preferable and more common to use an array in which only one doublet is excited directly and which contains a reflector system consisting of another doublet or a reflecting surface which is excited parasitically.

Dipole feeds although useful in practice have poor polarization characteristics and although natural when a coaxial antenna feed line is used are less convenient when the feed line is a waveguide. Since waveguides are more common in the microwave radar bands it is to be expected that waveguide feeds would be preferred in the majority of paraboloidal antennas.

The most easily constructed waveguide feed is simply an open ended waveguide. It is easy to permit a standard round or rectangular waveguide transmitting the dominant mode to radiate out into space toward the paraboloid. It will do this naturally with desirable phase, polarization and amplitude characteristics. It is purely coincidental, however, when this results in optimum amplitude characteristics. It is usually necessary to obtain these by tapering the feed line to form a waveguide aperture of the required size and shape. The aperture required may be smaller than a standard waveguide cross section so that its directivity will be less. In this case it may be necessary to 'load' it with dielectric material so that the power can be transmitted. It may be greater, in which case it is sometimes called an 'electromagnetic horn'. It may be greater in one dimension and less in the other, as when a paraboloidal section of large aspect ratio is to be illuminated.

If a single open ended waveguide or electromagnetic horn is used to feed a section of the paraboloid which includes the vertex, the waveguide feed line must partially block the reflected wave in order to be connected to the feed. To avoid this difficulty several rear waveguide feeds have been used. In this type of feed the waveguide passes through the vertex of the paraboloid and serves to support the feed at the focus. The energy can be caused to radiate back towards the reflector in any one of several ways, some of which involve reflecting rings or plates or parasitically excited doublets. The 'Cutler' feed[8] is perhaps the most successful and common rear feed. It operates by radiating the energy back towards the paraboloid through two apertures located and excited as shown in Fig. 23.

[8] C. C. Cutler, Loc. Cit.

8.5 *Parabolic Cylinders between Parallel Plates*

In 8.0 we saw that parabolic cylinders may be illuminated by line sources or that they may be confined between parallel plates and illuminated by point sources to produce line sources. In either of these two cases the characteristics which the feed should have are specified accurately by the conditions stated at the beginning of 8.4 for paraboloidal feeds with the exceptions that condition *c* must be reworded so that it applies to cylindrical rather than to spherical optics.

We will first consider parabolic cylinders bounded by parallel plates because in doing so we describe in passing one form of feed for unbounded parabolic cylinders. Two forms of transmission between parallel plates are used in practice.

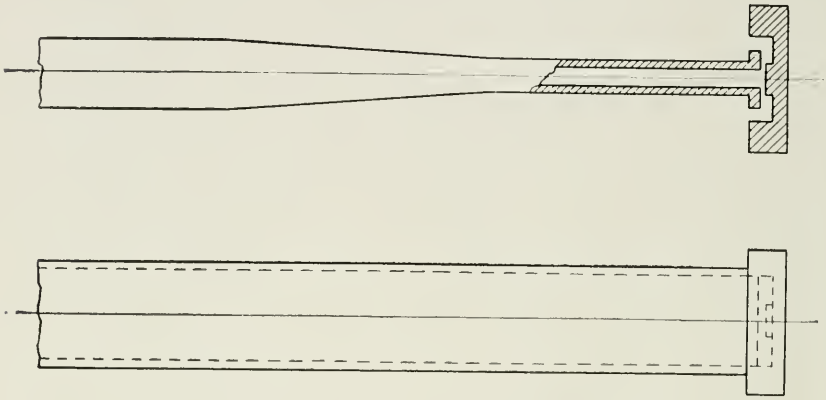


Fig. 23—Dual Aperture Rear Feed Horn.

a. The transverse electromagnetic (TEM) mode in which the electric vector is perpendicular to the plates. This is simply a slice of the familiar free space wave and can be propagated regardless of the spacing between the plates. It is the only mode that can travel between the plates if they are separated less than half a wavelength. Its velocity of propagation is independent of plate spacing.

b. The TE_{01} mode in which the electric vector is parallel to the plates. This mode is similar to the dominant mode in a rectangular waveguide and differs from it only in that it is not bounded by planes perpendicular to the electric vector. It can be transmitted only if the plate spacing is greater than half a wavelength, is the *only* parallel mode that can exist if the spacing is under a wavelength and is the only symmetrical parallel mode that can exist if the plate spacing is under three

halves of a wavelength. Its phase velocity is determined by the plate spacing in a manner given by the familiar waveguide formula

$$V_g = \frac{c}{\sqrt{\epsilon - \left(\frac{\lambda}{2a}\right)^2}}$$

where 'c' is the velocity of light, ϵ is the dielectric constant relative to free space of the medium between the plates, λ is the wavelength in air and 'a' is the plate spacing.

The TEM mode between parallel plates can be generated by extending the central conductor of a coaxial perpendicularly into or through the wave space and backing it up with a reflecting cylinder as indicated in Fig. 24.

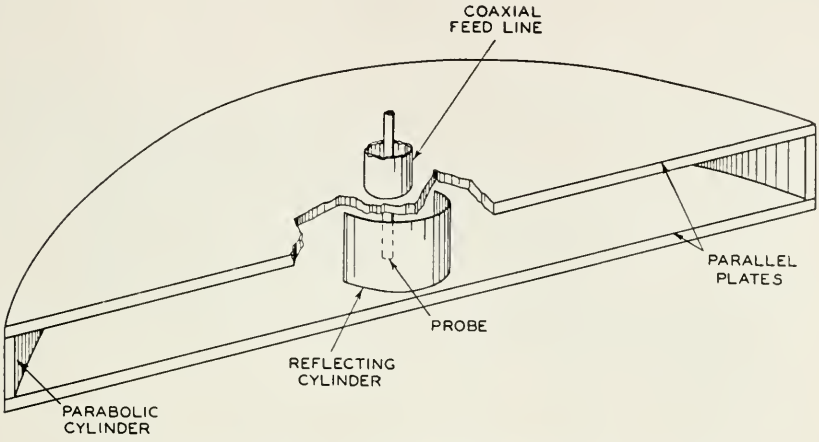


Fig. 24—Parabolic Cylinder Bounded by Parallel Plates. Probe Feed.

Alternatively this mode can be generated as indicated by Fig. 25 by a waveguide aperture with the proper polarization.

The TE_{01} mode, when used, is usually generated by a rectangular waveguide aperture set between the plates with proper polarization as indicated in Fig. 25. Care must be taken that only the desired mode is produced. The TEM mode will be unexcited if only the desired polarization is present in the feed. The next parallel mode is unsymmetrical and therefore even if it can be transmitted will be unexcited if the feed is placed symmetrically with respect to the two plates.

Parallel plate antennas as shown in figures 24 and 25 are useful where particularly large aspect ratios are required. The aperture dimension perpendicular to the plates is equal to the plate spacing and therefore small.

It can be increased somewhat by the addition of flares. The other dimension can easily be made large.

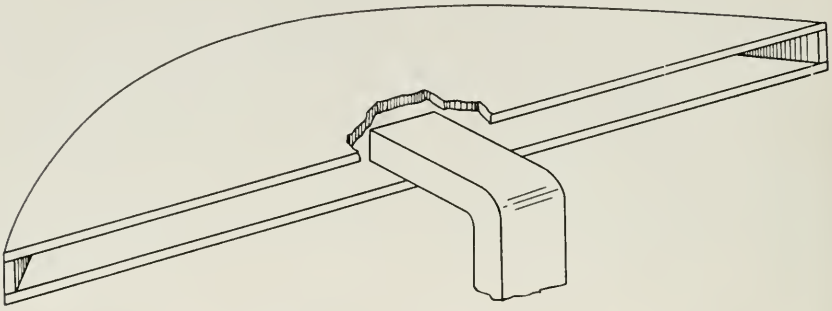


Fig. 25—Parabolic Cylinder Bounded by Parallel Plates. Wave Guide Feed.

Fig. 26.—Experimental 7′ x 32′ Antenna.

8.6 *Line Sources for Parabolic Cylinders*

A line source for a parabolic cylinder is physically an antenna with a long narrow aperture. Any means for obtaining such an aperture can be used in producing a line source. Parallel plate systems as described in 8.5 have been used as line sources in several radar antennas. The large (7′ x 32′)

experimental antenna shown in Fig. 26 was one of the first to illustrate the practicality of this design.

The horizontal pattern of the 7′ x 32′ antenna is plotted in Fig. 27. The horizontal beam width is seen to be of the order of 0.7 degrees.

The antenna illustrated in Fig. 26 is interesting in another way for it is a good example of a type of experimental construction which was extremely useful in wartime antenna development. Research and development engi-

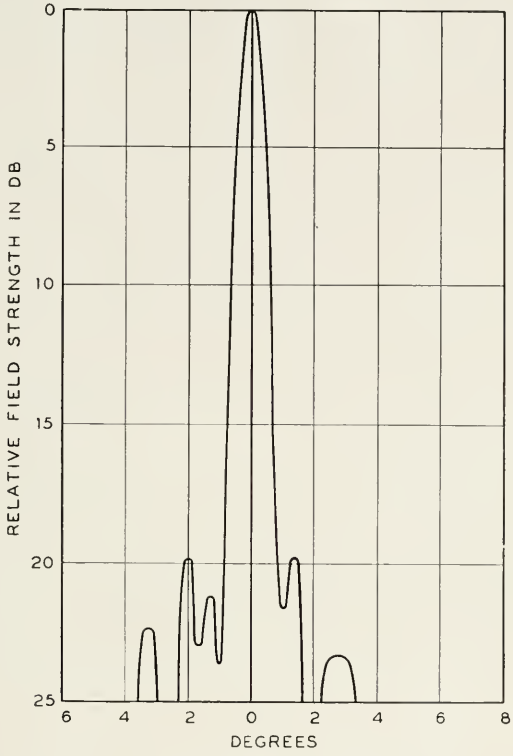


Fig. 27—7′ x 32′ Antenna, Horizontal Pattern.

neers found that they could often save months by constructing initial models of wood. Upon completion of a wooden model electrically important surfaces were covered with metal foil or were sprayed or painted with metal. Thus, where tolerances permitted, the carpenter shop could replace the relatively slow machine shop.

Another form of parallel plate line feed results when a plastic lens is placed between parallel plates and used as the focussing element. A linear array

of elements excited with the proper phase and amplitude can also be used. Some discussion of alternative approaches will appear in the section on scanning techniques.

8.7 *Tolerances in Parabolic Antennas*

The question of tolerances will always arise in practice. Ideal dimensions are only approximated, never reached. The ease of obtaining the required accuracy is an important engineering factor.

The tolerances in paraboloidal antennas or in parabolic cylinders illuminated by line sources can be divided into three general classes:

(a) Tolerances on reflecting surfaces.

(b) Tolerances on spacial relationships of feed and reflector.

(c) Tolerances on the feed.

When the actual reflector departs from the ideal parabolic curve deviations in the phase will result. These will tend to reduce the gain and increase the minor lobes. The effects of such deviations on the gain can be estimated with the help of Fig. 17. We should recall that an error of σ in the reflector surface will produce an error of about 2σ in the phase front. Based on this kind of argument and on experience reflector tolerances are generally set in practice to about $\pm \frac{\lambda}{16}$ or $\pm \frac{\lambda}{32}$ depending on the amount of beam deterioration that can be permitted.

In Fig. 28 are compared some electrical characteristics of two paraboloidal antennas, one employing a precisely constructed paraboloidal searchlight mirror and the other a carefully constructed wooden paraboloidal reflector with the same nominal contour. This comparison is revealing for it shows the harm that can be done even by small defects in the reflector surface. Although the two patterns are almost identical in the vicinity of the main beam, the general minor lobe level of the wooden reflector remains higher at large angles and its gain is less.

It must not, however, be assumed that a solid reflecting surface is necessary to insure excellent results. Any reflecting surface which reflects all or most of the power is satisfactory provided that it is properly located. Perforated reflectors, reflectors of woven material and reflectors consisting of gratings with less than half wavelength spacing are commonly used in radar antenna practice. These reflectors tend to reduce weight, wind or water resistance and visibility. Many of them will be described in Part III of this paper.

The feed of a parabolic reflector should be located so that its phase center coincides with the focus of the reflector. If it is located at an incorrect dis-

tance from the vertex a circular curvature of phase results and the system is said to be 'defocussed' (Sec. 3.9). As the feed is moved off the axis of the reflector the first effect is a shifting of the beam due to a linear variation of the phase (Sec. 3.8). For greater distances off axis a cubic component of phase error becomes effective (Sec. 3.10). Phase error, whether circular, cubic or more complex, results in a reduction in gain and usually in an increase of minor lobes. Although the effects of given amounts of phase curvature on the radiation characteristics of an antenna can be estimated by theoretical means, it is usually easier and quicker to find them experimentally.

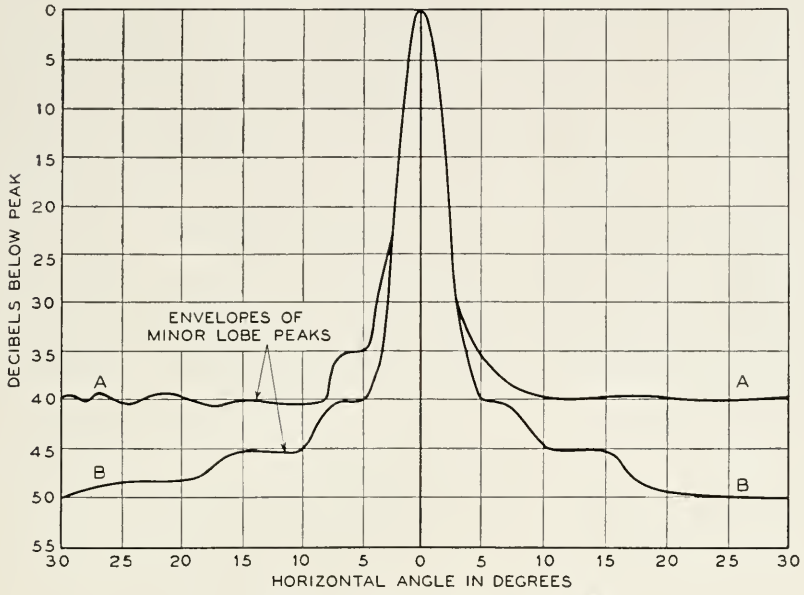


Fig. 28—Effect of Small Inaccuracies in Reflector.

The tolerances on the feed itself appear in various forms, many of which can be examined with the aid of transmission line theory and most of which are too detailed for discussion in this paper. It is generally true here also that experiment is a more effective guide than theory.

Experience has shown that when parallel plate systems are used, either as complete antennas or as line feeds for other elements, tolerances on the parallel conducting plates must be considered carefully. It is obvious that when the TE_{01} mode is used the plate spacing must be held closely, since the phase velocity is related to the spacing. This spacing can be controlled through the use of metallic spacers perpendicular to the plates. These

spacers, if small enough in cross section, do not disturb things unduly. The velocity of the TEM mode is, on the other hand, almost independent of the plate spacing. This mode is, however, more likely to cause trouble by leaks through joints and cracks in the plates.

9. Metal Plate Lenses

At visible wavelengths lenses have, in the past, been far more common than in the microwave region, due chiefly to the absence of satisfactory lens materials. A solid lens of glass or plastic with a diameter of several feet is a massive and unwieldy object. By zoning, which will be discussed below, these difficulties can be reduced but they still remain.

A new lens technique, particularly effective in the microwave region was developed by the Bell Laboratories during the war.[9] It is evident that any material in which the phase velocity is different from that of free space can be used to make a phase correcting lens. The material which is used in this new technique is essentially a stack of equally spaced metal plates parallel to the electric vector of the wave front and to the direction of propagation. Lenses made from this material are called 'Metal Plate Lenses'.

When the spacing between neighboring plates is between $\lambda/2$ and λ only one mode with electric vector parallel to the plates can be transmitted. This is the TE_{01} mode for which the phase velocity is given in Sec. 8.5. When the medium between the plates is air this equation can be converted into the expression

$$N = \sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}$$

for the effective index of refraction. Here λ is the wavelength in air and a is the plate spacing.

As a varies between $\lambda/2$ and λ, N varies as indicated in Fig. 29. In the neighborhood of $a = \lambda$, N is not far from 1 and as a approaches $\lambda/2$, N approaches 0. Since N is always less than 1 we see that there is an essential difference between metal plate lenses and glass or plastic lenses for which N is always greater than 1. This difference is seen in the fact that a glass lens corrects phases by slowing down a travelling wave front, while a metal lens operates in the reverse direction by speeding it up. This means that a convergent lens with a real focus must be thinner in the center than the edge, the opposite of a convergent optical lens (Fig. 30).

Unless the value of N is considerably different from 1 it is evident that very thick lens sections must be used to produce useful phase corrections. For this reason values of 'a' not far from $\lambda/2$ should be chosen. On the other hand values of 'a' too close to $\lambda/2$ would cause undesirably large reflections

[9] W. E. Kock, "Metal Lens Antennas", Proc. I. R. E., Nov., 1946.

from the lens surfaces and impose severe restrictions on the accuracy of plate spacings. The compromises that have been used in practice are $N = 0.5$ for which $a = 0.577\lambda$ and $N = 0.6$ for which $a = 0.625\lambda$.

Even with $N = 0.5$ or 0.6 lenses become thick unless inconveniently long focal distances are used. Thick lenses are undesirable not only because they occupy more space and are heavier but also because the plate spacing must be held to a higher degree of accuracy if the phase correction is to be as

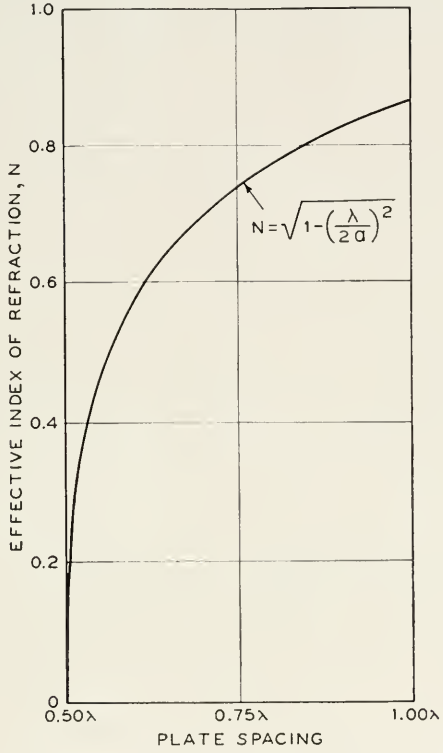


Fig. 29—Variation of Effective Index of Refraction with Plate Spacing in a Metal Plate Lens.

required. To get around these difficulties the technique of zoning is used.

Zoning makes use of the fact that if the phase of an electromagnetic vector is increased or decreased by any number of complete cycles the effect of the vector is unchanged. When applied to a metal plate lens antenna this means simply that wherever the phase correction due to a portion of the lens is greater than a wavelength this correction can be reduced by some integral number of wavelengths such that the residual phase correction is under one wavelength. If this is done it is evident that no portion of the

lens needs to correct the phase by more than one wavelength. It follows that no portion of the lens need to be thicker than $\lambda/(1 - N)$.

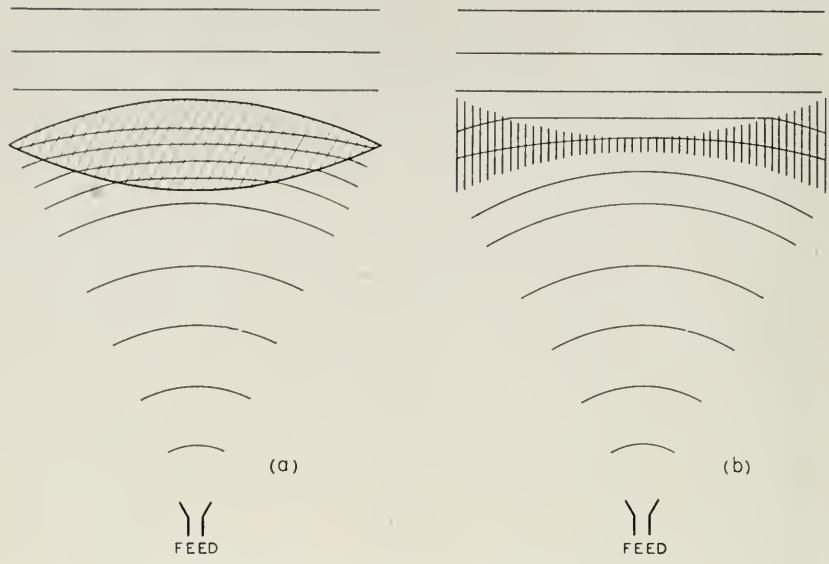


Fig. 30—Comparison of Dielectric and Metal Plate Lenses.

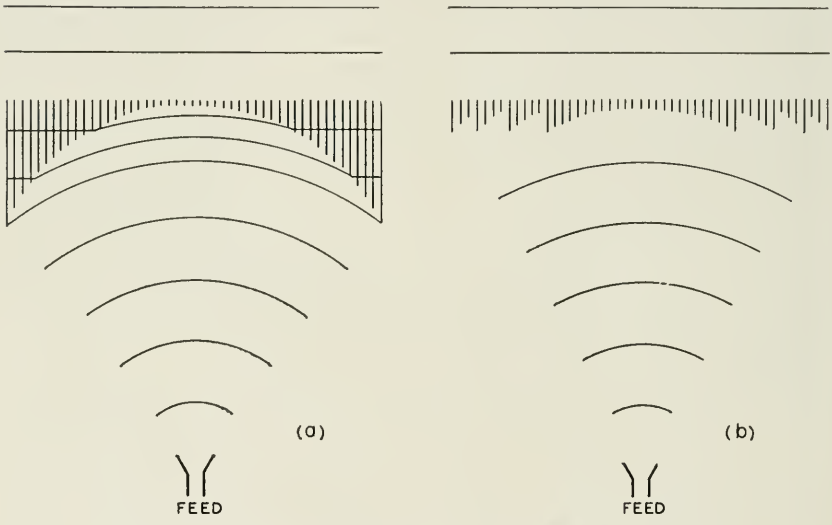


Fig. 31—Comparison of Unzoned and Zoned Metal Plate Lenses.

A cross section of a typical metal plate lens before and after zoning is illustrated in Fig. 31. This figure shows clearly why zoning reduces considerably the size and mass of a lens.

Zoning is not without disadvantages. One disadvantage is obviously that a zoned lens which is designed for one frequency will not necessarily work well at other frequencies. It is in principle possible to design a broad band zoned metal plate lens corresponding to the color compensated lenses used in good cameras. So far, however, this has not been necessary since band characteristics of simple lenses have been adequate.

Another difficulty that zoning introduces is due to the boundary regions between the zones. The wave front in this region is influenced partly by one zone and partly by the other and may as a result have undesirable phase and amplitude characteristics. This becomes serious only if especially short focal distances are used.

9.1 *Lens Antenna Configurations*

Any of the configurations which are possible with parabolic reflectors have their analogues when metal plate lenses are used. Circular lenses illuminated by point sources and cylindrical lenses illuminated by line sources are not only theoretically possible but have been built and used. Since a lens has two surfaces there is actually somewhat more freedom in lens design than in reflector design. Metal Plate Lenses have usually been designed with one surface flat, but the possibility of controlling both surfaces is emerging as a useful design factor where special requirements must be met.

Feeds for lenses should fulfill most of the same requirements as feeds for reflectors. We find a difference in size in lens feeds in that they must generally be more directive because of greater ratios of focal length to aperture. A difference in kind occurs because the feed is located behind the lens where none of the focussed energy can enter the feed or be disturbed by it. As a result some matching and pattern problems which arise in parabolic antennas are automatically absent when lenses are used.

In choosing a design for a lens antenna system with a given aperture one must compromise between the large size which is necessary when a long focal length is used and the more zones which result if the focal length is made short. Most metal plate lenses so far constructed have had focal lengths somewhere between 0.5 and 1.0 times the greatest aperture dimension.

9.2 *Tolerances in Metal Plate Lenses*

It is not difficult to see that phase errors resulting from small displacements or distortions of a metal plate lens are much less serious than those due to comparable distortions of a reflector surface. This follows from the fact that the lens operates on a wave which passes through it. If a portion of the lens is displaced slightly in the direction of propagation it is still operating on roughly the same portion of the wave front and gives it the same phase correction. If a portion of a reflector were displaced in the same way the error in the wave front would be of the order of twice the

displacement. Quantitative arguments show that less severe tolerances apply to all major structural dimensions of a metal lens antenna.

It is true of course that the dimensions of individual portions of the metal lens must be held with some accuracy. The metal plate spacing determines the effective index of refraction of the lens material. Where $N = 0.5$ it is customary to require that this be held to $\pm\lambda/75$, and where $N = 0.6$ to $\pm\lambda/50$. The thickness of the lens in a given region is less critical, and must be held to $\pm \dfrac{\lambda}{16\,(1 - N)}$ where it is desired to hold the phase front to $\pm\lambda/16$.

Fig. 32 illustrates clearly the drastic way in which the location of a lens can be altered without seriously affecting the pattern. It shows, incidentally, how a lens may behave well when used as the focussing element in a moving feed scanning antenna.

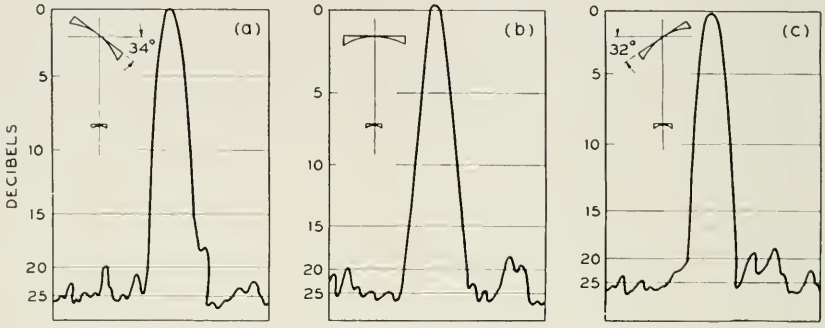


Fig. 32—Effect on Pattern of Lens Tilting.

9.3 *Advantages of Metal Plate Lenses*

On the basis of the above discussion we can see that metal plate lenses have certain considerable advantages. The most important of these is perhaps found in the practical matter of tolerances. It is a comparatively simple matter to hold dimensions of small objects to close tolerances but quite another thing to hold dimensions of large objects closely under the conditions of modern warfare. This advantage emerges with increasing importance as the wavelength is reduced.

Metal plate lenses have contributed a great degree of flexibility to radar antenna art. When they are used two surfaces rather than one may be controlled, and the dielectric constant can be varied within wide limits. Independent control in the two polarizations may be applied. We can confidently expect that they will become increasingly popular in the radar field.

10. Cosecant Antennas

One of the earliest uses of radar was for early warning against aircraft.

The skies were searched for possible attackers with antennas which rotated continuously in azimuth. An equally important but later use appeared with the advent of great bombing attacks. Bombing radars 'painted' maps of the ground which permitted navigation and bombing during night and under even the worst weather conditions. In these radars also the antennas were rotated in azimuth, either continuously through 360° or back and forth through sectors.

The majority of radars designed to perform these functions provided vertical coverage by means of a special vertical pattern rather than a vertical scan. It can easily be seen that such a pattern would have to be 'special.' If we assume, for example, that a bombing plane is flying at an altitude of 10,000 feet, then the radar range must be 10,000 *csc* 60° = 11,500 feet if a target on the ground at a bomb release angle of 60° from the horizontal is to be seen. Such a range would by no means be enough to pick up the target at say 10° in time to prepare for bombing, for then a range of 10,000 csc 10° = 57,600 feet would be required. This range is far more than is necessary for the 60° angle. It appears then that in the most efficient design the radar range and therefore the radar antenna gain, must be different in different directions.

The required variation of gain with vertical direction could be specified in any one of several ways. It seems natural to specify that a given ground target should produce a constant signal as the plane flies towards it at a constant altitude. Neglecting the directivity of the target this will occur if the amplitude response of the antenna is given by $E = E_0 \csc\theta$. This same condition will apply by reciprocity to an early warning radar antenna on the ground which is required to obtain the same response at all ranges from a plane which is flying in at a constant altitude.

This condition is not alone sufficient to specify completely the vertical pattern of an antenna. For one thing it can obviously not be followed when $\theta = 0$, for this would require infinite gain in this direction. Therefore a lower limit to the value of θ for which the condition is valid must be set. In addition an upper limit less than 90° is specified whenever requirements permit, since control at high angles is especially difficult. When the limits have been set it still remains to specify the magnitude of the constant E_0. This can be done by specifying the range in one particular direction. This specification must of course be consistent with all the factors that determine gain, including the reduction due to the required vertical spread of the pattern.

10.1 *Cosecant Antennas based on the Paraboloid*

It is evident that the standard paraboloidal antennas so far discussed will not produce cosecant patterns. These patterns being unsymmetrical will result only if the wave front phase and amplitude are especially controlled.

On the other hand, because paraboloidal antennas are simple and common it is natural that many cosecant designs should be based on them. These designs can be classified into two groups, those in which the reflector is modified and those in which the feed is modified.

Some early cosecant antennas were made by introducing discontinuities in paraboloidal reflectors as illustrated in Fig. 33. These controlled the radiation more or less as desired over the desired cosecant pattern but pro-

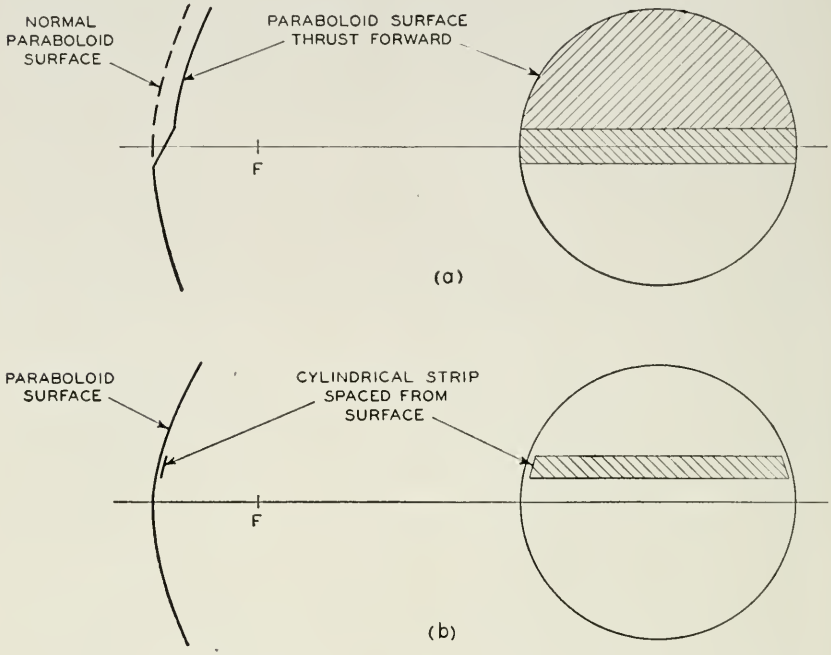


Fig. 33—Some Cosecant Antennas Based on the Paraboloid (Cosecant Energy Downward).

duced rather serious minor lobes elsewhere. This difficulty can be overcome through the use of a continuously distorted surface as illustrated in Fig. 34. This reflector, first used at the Radiation Laboratories, is a normal paraboloid in the lower part whereas the upper part is the surface that would be obtained by rotating the parabola through the vertex of the upper part about its focal point.

Several types of feed have been used in combination with paraboloids to produce cosecant patterns. These are usually arrays which operate on the principle that each element is a feed which contributes principally to one

region of the vertical pattern. The elements may be dipoles or waveguide apertures fed directly through the antenna feed line or they may be reflectors which reradiate reflected energy originating from a single primary source. No matter how excited they must be properly controlled in phase, amplitude and directivity.

Cosecant antennas based on the paraboloid are common and can sometimes fulfill all requirements with complete satisfaction. Nevertheless they

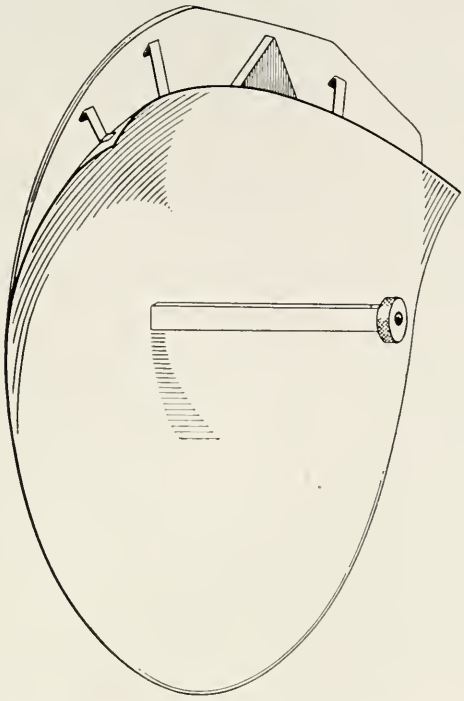


Fig. 34—Barrel Cosecant Antenna (Cosecant Energy Downward).

suffer from certain disadvantages. The most serious of these is that they lack resolution at high vertical angles, that is the beam is wider horizontally at high angles. This is to be expected for reasons of phase alone, for a paraboloidal reflector is, after all, designed to focus in only one direction. If phase difficulties were completely absent however, azimuthal resolution at high angles would still be destroyed because of cross polarized components of radiation. These components arise naturally from doubly curved reflectors, even simple paraboloids. They are sometimes overlooked when antennas are measured in a one way circuit with a linearly polarized test field, but must obviously be considered in radar antennas.

10.2 *Cylindrical Cosecant Antennas*

Harmful cross polarized radiation is produced by doubly curved reflectors. This radiation is difficult to control and therefore undesirable where a closely controlled cosecant characteristic at high angles is required. Although not at first evident, it seems natural now to bypass polarization difficulties through the use of singly curved cylindrical reflectors. These reflectors if illuminated with a line source of closely controlled linear polarization provide a beam which is linearly polarized. This beam has also in azimuth approximately the directivity of the line source at all vertical angles. It is thus superior in two significant respects to cosecant beams produced by doubly curved reflecting surfaces.

A cylindrical cosecant antenna consists of a cylindrical reflector illuminated by a line source. Part of the cylinder is almost parabolic and contributes chiefly to the strong part of the beam which lies closest to the horizontal. This part is merged continuously into a region which departs considerably from the parabolic and contributes chiefly to the radiation at higher angles.

Although wave front principles can be used and certainly must not be violated, the principles of geometrical optics have been particularly effective in the determination of cosecant reflector shapes. The detailed application of these principles will not be discussed here. While applying the geometrical principles the designer must be sure that the over-all size and configuration of the antenna can produce the results he wants. He must design a line source with the desired polarization and horizontal pattern and a vertical pattern which fits in with the cosecant design. In addition he must take particular care to reduce sources of pattern distortion to a level at which they cannot interfere significantly with the lowest level of the cosecant 'tail'.

11. Lobing

In many of the tactical situations of modern war radar can be used to provide fire control information. Radar by its nature determines range and microwave radar with its narrow well defined beams is a natural instrument for finding directions to a target, whether the missile to be sent to that target is a shell, a torpedo or a bomb. In fire control radar, as opposed to search or navigational radar, two properties of the antenna deserve particular attention. These are the *accuracy* and the *rate* with which direction to a target can be measured.

Lobing is a means which utilizes to the fullest extent the accuracy available from a given antenna aperture and which increases, usually as far as is desired, the rate at which this information is provided and corrected.

A lobing antenna which is to provide information concerning one angle only, azimuth for example, is capable of producing two beams, one at a time, and of switching rapidly from one to the other. This process is called *Lobe Switching.* The two beams are nearly coincident, differing in direction by about one beam width. When the signals from the two beams are compared, they will be equal only if the target lies on the bisector between the beams (Fig. 35). The two signals can be compared visually on an indicator screen of the radar or they can be compared electrically and fed directly into circuits which control the direction of fire.

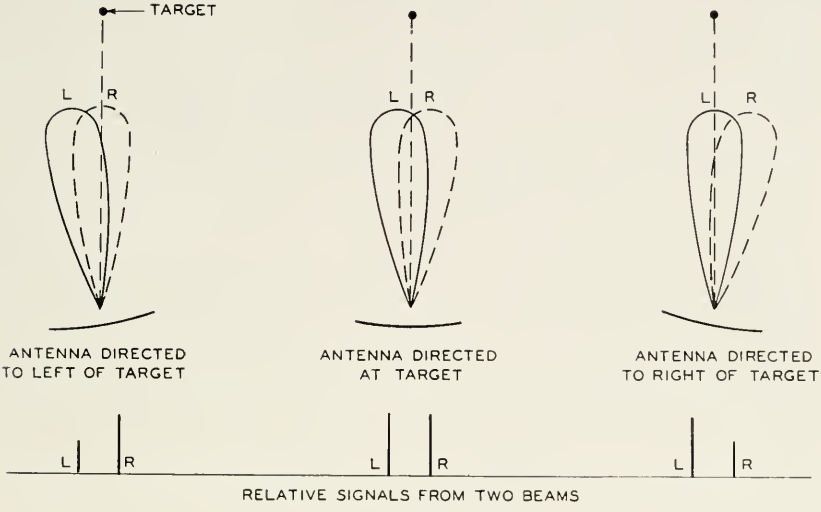


Fig. 35—Lobe Switching.

When two perpendicular directions are to be determined, such as the elevation and azimuth required by an anti-aircraft battery, four or in principle three discrete beams can be used. Radar antennas designed for solid angle coverage more commonly, however, produce a single beam which rotates rapidly and continuously around a small cone. This rotation is known as *conical lobing.* A comparison of amplitudes in a vertical plane can then be used to give the elevation of the target and a similar comparison in a horizontal plane to give its azimuth. Here too the electrical signals can be compared visually on an indicator screen, but an electrical comparison will provide continuous data which can be used to aim the guns and at the same time to cause the radar antenna to follow the target automatically.

11.1 *Lobe Switching*

Two methods of lobe switching are common. In one of these the lobing antenna is an array of two equally excited elements. Each of these ele-

ments occupies one half of the final antenna aperture, and provides a uniphase front across this half. If the two elements were excited with the same phase the radiation maximum of the resulting antenna beam would occur in a direction at right angles to the combined phase front. If the phase of one element is made to lag behind that of the other by a small amount, 60° say, the phase of the combined aperture will of course be discontinuous with a step in the middle. This discontinuous phase front will approximate with a small error, a uniphase wave front which is tilted somewhat with respect to the wave fronts of the elements. The phase shift will therefore result in a slight shift of the beam away from the normal direction. When the phase shift is reversed the beam shift will be reversed. Two properly designed elementary antennas in combination with a means for rapidly changing the phase will therefore constitute a lobe switching antenna. Such an antenna is described more in detail in Sec. 14.6.

Another method of lobe switching is more natural for antennas based on optical principles. In this method two identical feeds are placed side by side in the focal region of the reflector. When one of these feeds illuminates the reflector a beam is produced which is slightly off the normal axial direction. Illumination by the other feed produces a second beam which is equally displaced in the opposite direction. The lobe of the antenna switches rapidly when the two feeds are activated alternately in rapid succession. The antenna must use some form of rapid switching appropriate to the antenna feed line. In several applications switches are used which depend on the rapid tuning and detuning of resonant cavities or irises.

11.2 *Conical Lobing*

A conically lobing antenna produces a beam which nutates rapidly about a fixed axial direction. This is usually accomplished by rotating or nutating an antenna feed in a small circle about the focus in the focal plane of a paraboloid or lens. This antenna feed can be a spinning asymmetrical dipole or a rotating or nutating waveguide aperture. It can result in a beam with linear polarization which rotates as the feed rotates, or preferably in a beam for which the polarization remains parallel to a fixed direction. The beam itself must be nearly circularly symmetric so that the radar response from a target in the axial direction will not vary with the lobing. The reflector or lens aperture is consequently usually circular.

When the antenna is small it is sometimes easier to leave the feed fixed and to produce the lobing by moving the reflector.

12. RAPID SCANNING

A lobing radar can provide range and angular information concerning a single target rapidly and accurately but these things are not always enough.

It is sometimes necessary to obtain accurate and rapid information from all regions within an agular sector. It may be necessary to watch a certain region of space almost continuously in order to be sure of picking up fast moving targets such as planes. To accomplish any of these ends we must use a *rapid scanning* radar. A rapid scanning radar antenna produces a beam which scans continuously through an angular sector. The beam may sweep in azimuth or elevation alone or it may sweep in both directions to cover a solid angle. An azimuth or elevation scan may be sinusoidal or it may occur linearly and repeat in a sawtooth fashion. Solid angle scanning may follow a spiral or flower leaf pattern or it might be a combination of two one way scans. A combination of scanning in one direction and lobing in the other is sometimes used.

Scanning antennas must, unfortunately, be constructed in obedience to the same principles which regulate ordinary antennas. The same attention to phase, amplitude, polarization and losses is necessary if comparable results are to be obtained. When scanning requirements are added to these ordinary ones new problems are created and old ones made more difficult.

An antenna in order to scan in any specified manner must act to produce a wave front which has a constant phase in a plane which is always normal to the required beam direction. This can be done in several different ways. The simplest of these, electrically, is to rotate a fixed beam antenna as a whole in the required fashion. This can be called *mechanical scanning*. Alternatively an antenna array can be scanned if it is made up of suitable elements and the relative phases of these elements can be varied appropriately. This can be called *array scanning*. Thirdly, *optical scanning* can be produced by moving either the feed or the focussing element of a suitably designed optical antenna.

12.1 *Mechanical Scanning*

Electrical complexities of other types of rapid scanners are such that it is probably not going too far to say that the required scan should be accomplished by mechanical means wherever it is at all practical. This applies to radar antenna scans which occur at a slow or medium rate. Search antennas, whether they rotate continuously through 360° or back and forth over a sector are scanners in a sense but the scan is usually slow enough to be performed by rotating the antenna structure as a whole. As the scan becomes more rapid, mechanical problems become more severe and electrically scanning antennas appear more attractive.

Mechanical ingenuity has during the war extended the range in which mechanical scanners are used. One important and eminently practical mechanical rapid scanner, the 'rocking horse' is now in common use (Fig. 36). This antenna is electrically a paraboloid of elliptical aperture illu-

minated by a horn feed, a combination which produces excellent electrical characteristics. The paraboloid and feed combination is made structurally strong and is pivoted to permit rotational oscillation in a horizontal plane. It is forced to oscillate by a rigid crank rod which is in turn driven by an eccentric crank on a shaft. The shaft is belt driven by an electric motor and its rotational rate is held nearly constant by a flywheel. The mechanical arrangement described so far would oscillate rotationally in an approximately sinusoidal fashion. Since every action has an equal and opposite reaction it would, however, react by producing an oscillatory torque on its

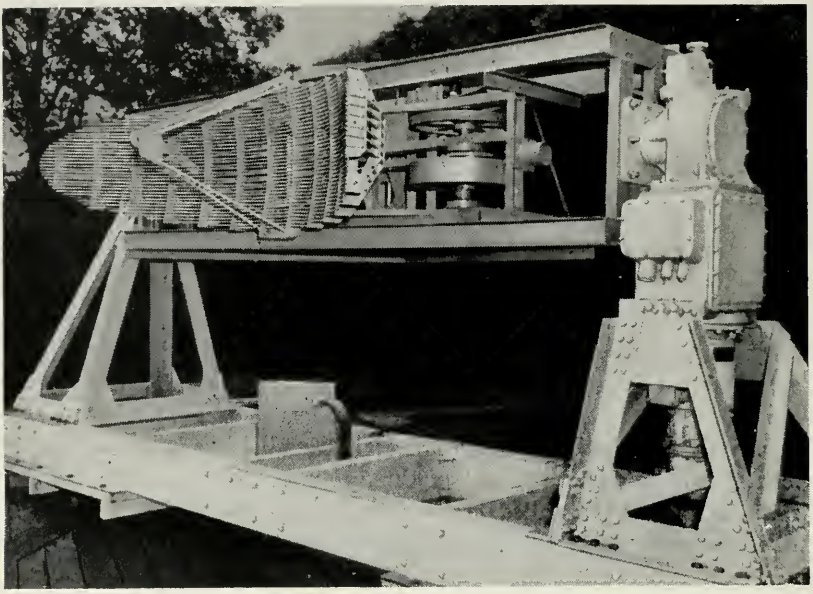


Fig. 36—Experimental Rocking Horse Antenna.

mounting. Since the antenna is large and the oscillation rapid this would produce a severe and undesirable vibration. To get around this difficulty an opposite and balancing rotating moment is introduced into the mechanical system. This appears in the form of a pivoted and weighted rod which is driven from the same eccentric crank by another and almost parallel crank arm.

Although not theoretically perfect the rotational 'dynamic' balancing described permits the antenna to scan without serious vibration. One form of this antenna will be described in a later section.

12.2 *Array Scanning*

During our discussion of general principles in Part II, we saw that an antenna wave front can be synthesized by assembling an array of radiating

elements and distributing power to it through an appropriate transmission line network. If the radiation characteristics of the array are to be as desired the electrical drive of each element must have a specified phase and amplitude. In addition each element must in itself have a satisfactory characteristic and the elements must have a proper spacial relationship to each other.

Such array antennas have been extremely useful in the 'short wave' bands where wavelengths and antenna sizes are many times larger than at most radar wavelengths but for fixed beam radar antennas they have been largely superceded by the simpler optical antennas. Where a rapidly scanning beam is desired, however, they possess certain advantages which were put to excellent use in the war. These advantages spring from the possibility of scanning the beam of an array through the introduction of rapidly varying phase changes in its transmission line distributing system.

Let us first examine certain basic conditions that must be fulfilled if an array antenna is to provide a satisfactory scan. The pattern of any array is merely the sum of the patterns of its elements taking due account of phase, amplitude and spacial relationships. If all elements are alike and are spaced equally along a straight line it is not difficult to show that a mathematical expression for the pattern can be obtained in the form of a product of a factor which gives the pattern of a single element and an array factor. The array factor is an expression for the pattern of an array of elements each of which radiates equally in all directions. Since each of the elements is fixed in direction it is only through control of the array factor that the scan can be obtained.

If we excite all points of a continuous aperture with equal phase and a smoothly tapered amplitude the aperture produces a beam with desirable characteristics at right angles to itself and no comparable radiation elsewhere. Similarly if we excite all elements of an array of identical equally spaced circularly radiating elements with equal phase and a smoothly tapered amplitude the array will produce a beam with desirable characteristics at right angles to itself. It will also produce a beam in any *other* direction for which waves from the elements can add up to produce a wave front. Such other directions will exist whenever the array spacing is greater than one wavelength.

In order to see this more clearly let us examine Fig. 37, where line XX' represents an array of elements. From each element to the line AA' is a constant distance, so AA' is obviously parallel to a wave front when the elements are excited with equal phase. If we can find a line BB' to which the distance from each element is exactly one wavelength more or less than from its immediate neighbors then it too is parallel to a wavefront, for energy reaching it from any element of the array will have the same phase

except for an integral number of cycles. The same will apply to a line CC', to which the distance from each element is exactly two wave lengths more or less than from its immediate neighbors, or to any other line where this difference is any integral number of wavelengths.

Now in no radar antenna do we desire two or more beams for they will result in loss of gain and probably in target confusion. The array must therefore be designed so that for all positions of scan all beams except one will be suppressed. This will automatically occur if the array spacing is somewhat less than one wavelength. If the array spacing is greater than one wavelength these extra beams will appear in the array factor; they

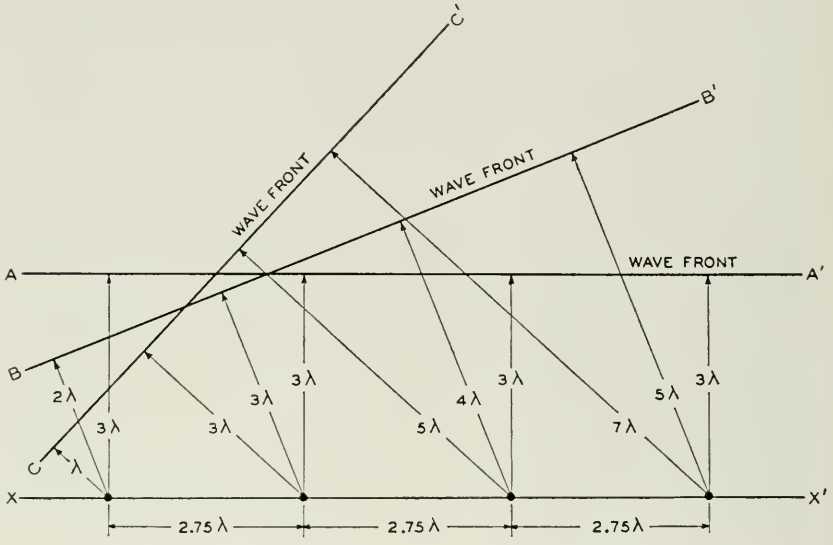


Fig. 37—Some Possible Wave Fronts of an Array of Elements Spaced 2.75 λ.

must therefore be suppressed by the pattern of a single element. The pattern of an element must in other words, have no significant components in any direction where an extra beam can occur.

Where elements with only side fire directivity are spaced more than a wavelength apart in a scanning array it is almost impossible to obtain adequate extra lobe suppression. If these elements are spaced by the minimum amount, that is by exactly the dimensions of their apertures and all radiate in phase they may indeed just manage to produce a desirable beam. A little analysis shows however that an appreciable phase variation from element to element, even though linear, will introduce a serious extra lobe. To get around this difficulty elements with some end fire directivity must be used.

A simple end fire element, and one that has been used in practice, is the 'polyrod' (Fig. 38). A polyrod, is as its name implies, a rod of polystyrene. This rod, if inserted into the open end of a waveguide, and if properly proportioned and tapered, will radiate energy entering from the waveguide from points which are distributed continuously along its length. If the

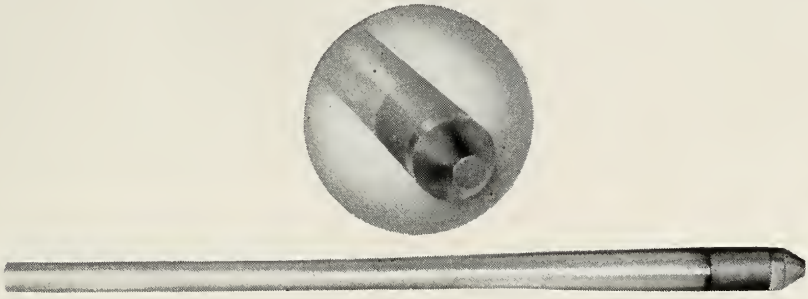


Fig. 38—A Polyrod.

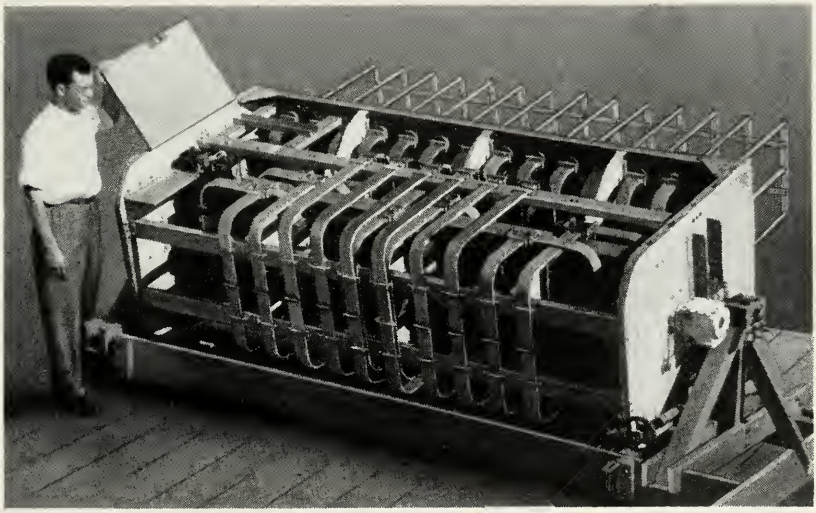


Fig. 39.—Experimental Polyrod Array.

wave in the polyrod travels approximately with free space velocity it will produce a radiation maximum in the direction of its axis. The radiation pattern of the polyrod will have a shape which is characteristic of end fire arrays, narrower and flatter topped than the pattern of a side fire array which occupies the same *lateral* dimension. This elementary pattern can be fitted in well with the array factor of a scanning array.

Such a scanning array is shown in Fig. 39 and will be described in

greater detail in section 14.8. Each element of this array consists of a fixed vertical array of three polyrods. This elementary array provides the required vertical pattern and has appropriate horizontal characteristics. Fourteen of these elements are arranged in a horizontal array with a spacing between neighbors of about two wavelengths. Energy is distributed among the elements with a system of branching waveguides. Thirteen rotary phase changers are inserted strategically in the distributing system. Each phase change is rotating continuously and shifts the phase linearly from 0° to 360° twice for each revolution. As the phase changers rotate the array produces a beam which sweeps repeatedly linearly and continuously across the scanning sector.

When elements of a scanning array are spaced considerably less than one wavelength it is a very simple matter to obtain a suitable elementary pattern, for the array factor itself has only a single beam. This advantage is offset by the greater number of elements and the consequent greater complexity of distributing and phase shifting equipment. In one useful type of scanning antenna however distributing and phase shifting is accomplished in a particularly simple manner. Here the distributing system is merely a waveguide which can transmit only the dominant mode. The wide dimension of the guide is varied to produce the phase shifts required for scanning. The elements are dipoles. The center conductor of each dipole protrudes just enough into the guide to pick up the required amount of energy.

It is evident from the above discussion that such a waveguide fed dipole array will produce a single beam in the normal direction only if the dipoles are all fed in phase and are spaced less than a wavelength. It is therefore not satisfactory to obtain constant phase excitation by tapping the dipoles into the guide at successive guide wavelengths for these are greater than free space wavelengths. Consequently the dipoles are tapped in at successive half wavelengths in the guide and reversed successively in polarity to compensate for the successive phase reversals due to their spacing.

This type of array provides a line source which can be scanned by moving the guide walls. In order to leave these mechanically free suitable wave trapping slots are provided along the length of the array.

A practical antenna of this type will be described in Sec. 16.3.

12.3 *Optical Scanning*

With a camera or telescope all parts of an angular sector or field are viewed simultaneously. We would like to do the same thing by radar means, but since this so far appears impossible we do the next best thing by looking at the parts of the field in rapid succession. Nevertheless certain points of similarity appear. These points are emphasized by a survey of the fixed

beam antenna field for there we find optical instruments in abundance, parabolic reflectors and even lenses.

It is not a very big step to proceed from an examination of optical systems to the suggestion that a scanning antenna can be provided by moving a feed over the focal plane of a reflector. Nevertheless experience shows that this will not be especially profitable unless done with due caution. The first effect of moving the feed away from the focus in the focal plane of a paraboloid is indeed a beam shift but before this process has gone far a third order curvature of the phase front is produced and is accompanied by a serious deterioration in the pattern and reduction in gain. This difficulty or aberration is well known in classical optical theory and is called coma. Coma is typified by patterns such as the one shown in Fig. 16. It is the first obstacle in the path of the engineer who wishes to design a good moving feed scanning antenna.

Coma is not an insuperable obstacle however. Its removal can be accomplished by the application of a very simple geometrical principle. This principle can be stated as follows: "The condition for the absence of coma is that each part of the focussing reflector or lens should be located on a circle with center at the focus."

This condition can be regarded as a statement of the spacial relationship required between the feed and all parts of the focussing element. It is a condition which insures that the phase front will remain nearly linear when the feed is moved in the focal plane. It can be applied approximately whether the focussing element is a reflector or a lens and to optical systems which scan in both directions as well as those which scan in one direction.

Coma is usually the most serious aberration to be reckoned with in a scanning optical system, but it is by no means the only one. Any defect in the phase and amplitude characteristic which arises when the feed is moved can cause trouble and must be eliminated or reduced until it is tolerable. Another defect in phase which arises is 'defocussing'. Defocussing is a square law curvature of phase and arises when the feed is placed at an improper distance from the reflector or lens. Its effect may be as shown in Fig. 14. It can in principle always be corrected by moving the feed in a correctly chosen arc, but this is not always consistent with other requirements on the system. In addition to troubles in phase an improper amplitude across the aperture of the antenna will arise when the feed is translated unless proper rotation accompanies this motion.

To combat the imperfections in an optical scanning system we can choose over-all dimensions in such a way that they will be lessened. Thus it is generally true that an increase in focal length or a decrease in aperture will increase the scanning capabilities of an optical system. This alone is usually not enough, however, we must also employ the degrees of free-

dom available to us in the designing of the focussing element and the feed motion to improve the performance. If the degrees of freedom are not enough we must, if we insist on an optical solution introduce more. This could in principle result in microwave lenses similar to the four and five element glass lenses found in good cameras, but such complication has not as yet been necessary in the radar antenna art.

Since military release has not been obtained as this article goes to press we must omit any detailed discussion of optically scanning radar antenna techniques.

PART III

MILITARY RADAR ANTENNAS DEVELOPED BY THE BELL LABORATORIES

13. General

In the final part of this paper we will describe in a brief fashion the end products of radar antenna technology, manufactured radar antennas. Without these final practical exhibits the foregoing discussion of principles and methods might appear academic. By including them we hope to illustrate in a concrete fashion the rather general discussion of Parts I and II.

The list of manufactured antennas will be limited in several ways. Severe but obviously essential are the limitations of military security. In addition we will restrict the list to antennas developed by the Bell Laboratories. In cases where invention or fundamental research was accomplished elsewhere due credit will be given. Finally the list will include only antennas manufactured by contract. This last limitation excludes many experimental antennas, some initiated by the Laboratories and some by the armed forces.

It is worthwhile to begin with an account of the processes by which these antennas were brought into production. The initiating force was of course military necessity. The initial human steps were taken sometimes by members of the armed forces who had definite needs in mind and sometimes by members of the Laboratories who had solutions to what they believed to be military needs.

With a definite job in mind conferences between military and Laboratories personnel were necessary. Some of these dealt with legal or financial matters, others were principally technical. In the technical conferences it was necessary at an early date to bring military requirements and technical possibilities in line.

As a result of the conferences a program of research and development was often undertaken by the Laboratories. An initial contract was signed which

called for the delivery of technical information, and sometimes for manufacturing drawings and one or more completed models. Usually the antenna was designed and manufactured as part of a complete radar system, sometimes the contract called for an antenna alone.

After preliminary work had been undertaken the status of the job was reviewed from time to time. If preliminary results and current military requirements warranted a manufacturing contract was eventually drawn up and signed by Western Electric and the contracting government agency. This contract called for delivery of manufactured radars or antennas according to a predetermined schedule.

Research and development groups of the Laboratories cooperated in war as in peace to solve technical problems and accomplish technical tasks. Under the pressure of war the two functions often overlapped and seemed to merge, yet the basic differences usually remained.

Members of the Research Department, working in New York and at the Deal and Holmdel Radio Laboratories in New Jersey were concerned chiefly with electrical design. It was their duty to understand fully electrical principles and to invent and develop improved methods of meeting military requirements. During the war it was usually their responsibility to prescribe on the basis of theory and experiment the electrical dimensions of each new radar antenna.

A new and difficult requirement presented to the Research Department was sometimes the cause of an almost personal competition between alternative schemes for meeting it. Some of these schemes were soon eliminated by their own weight, others were carried side by side far along the road to production. Even those that lost one race might reappear in another as a natural winner.

In the Development Groups working in New York and in the greatly expanded Whippany Radio Laboratory activity was directed towards coordination of all radar components, towards the establishment of a sound, well integrated mechanical and electrical design for each component and towards the tremendous task of preparing all information necessary for manufacture. It was the job of these groups also to help the manufacturer past the unavoidable snarls and bottlenecks which appeared in the first stages of production. In addition development personnel frequently tested early production models, sometimes in cooperation with the armed forces.

As we have intimated, research and development were indistinguishable at times during the war. Members of the research department often found themselves in factories and sometimes in aircraft and warships. Development personnel faced and solved research problems, and worked closely with research groups.

For several years when pressure was high the effort was intense; at times feverish. Judging by military results it was highly effective. Some of the material results of this effort are described in the following pages.

14. Naval Shipborne Radar Antennas

14.1 *The SE Antenna*[10]

Very early in the war, the Navy requested the design of a simple search radar system for small vessels, to be manufactured as quickly as possible in order to fill the gap between design and production of the more complex search systems then in process of development. The proposed system was to be small and simple, to permit its use on vessels which otherwise would be unable to carry radar equipment because of size or power supply capability. This class of vessel included PT boats and landing craft.

The antenna designed for the SE system is housed as shown in Fig. 40. It was adapted for mounting on the top or side of a small ship's mast, and is rotated in azimuth by a mechanical drive, hand operated. The paraboloid reflector is 42 inches wide, 20 inches high, and is illuminated by a circular aperture 2.9 inches in diameter. In the interests of simplicity, the polarization of the radiated beam was permitted to vary with rotation of the antenna.

The SE antenna was operated at 9.8 cm, and fed by $1\frac{1}{2}$ x 3 rectangular waveguide. At the antenna base, a taper section converted from the rectangular waveguide to 3″ round guide, through a rotating joint directly to the feed opening.

Characteristics of the SE antenna are given below:

Wavelength	9.7 to 10.3 cm	
Reflector	42″ W x 20″ H	
Gain	25 db	
Horizontal Beam Width	6°	
Vertical Beam Width	12°, varying somewhat with polarization	
Standing Wave	9.7to10.0 cm	4.0 db
	10.0 to 10.3 cm	6.0 db

14.2 *The SL Radar Antenna*[11]

The SL radar is a simple marine search radar developed by Bell Telephone Laboratories for the Bureau of Ships. During the war, over 1000 of these radars were produced by the Western Electric Company and installed on Navy vessels of various categories. The principal field for installation was destroyer escort craft ("DE"s). Figure 41 shows an SL antenna installation aboard a DE. The antenna is covered, for wind and

[10] Written by R. J. Phillips.
[11] Written by H. T. Budenbom.

weather protection, in a housing which can transmit 10 cm radiation. Visible also is the waveguide run down the mast to the r.f. unit.

The SL radar provides a simple non-stabilized PPI (Plan Position Indicator) display. The antenna is driven by a synchronous motor at 18 rpm. Horizontal polarization is used to minimize sea clutter. The

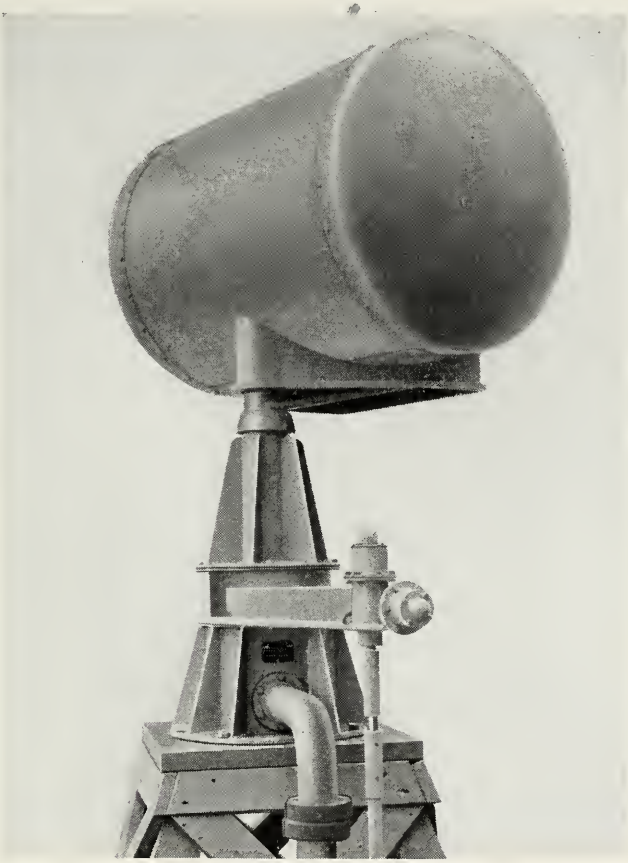


Fig. 40—SE Antenna.

radiating structure, shown in Figure 42, consists of a 20″ sector of a 42″ paraboloid. The resulting larger beam width in the vertical plane is provided in order to improve the stability of the pattern under conditions of ship roll. Figure 43 illustrates the path of the transmitted wave from the SL r.f. unit to the antenna. It also illustrates the manner in which horizontally polarized radiation is obtained. The diagram shows the position of

Fig. 41—SL Antenna Aboard DE.

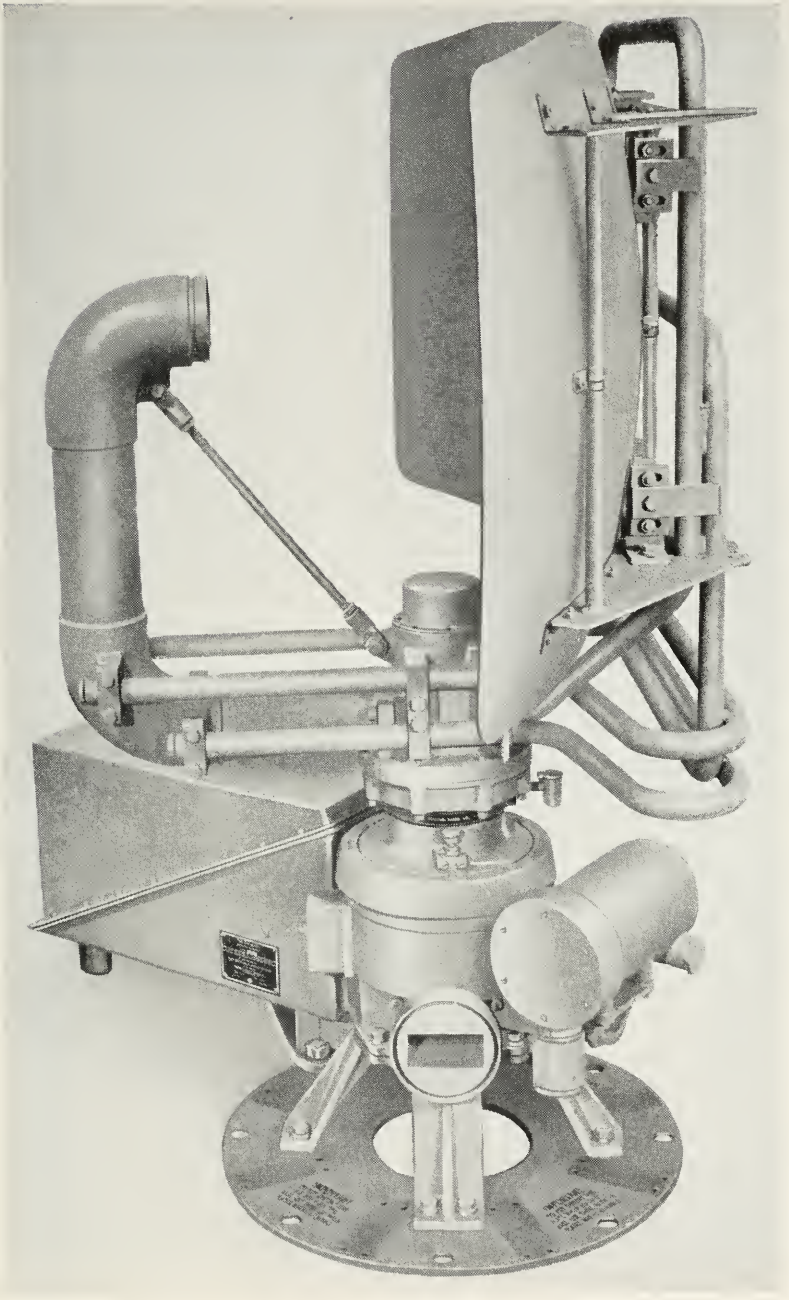


Fig. 42—SL Antenna.

the electric force vector in traversing the waveguide run. The path from the r.f. unit is in rectangular guide ($TE_{1,0}$ mode) through the right angle bend, to the base of the rotary joint. A transducer which forms the base portion of the joint converts to the TM_{01} mode in circular pipe. For this mode, the electric field has radial symmetry, much as though the waveguide were a coaxial line of vanishingly small inner conductor diameter.

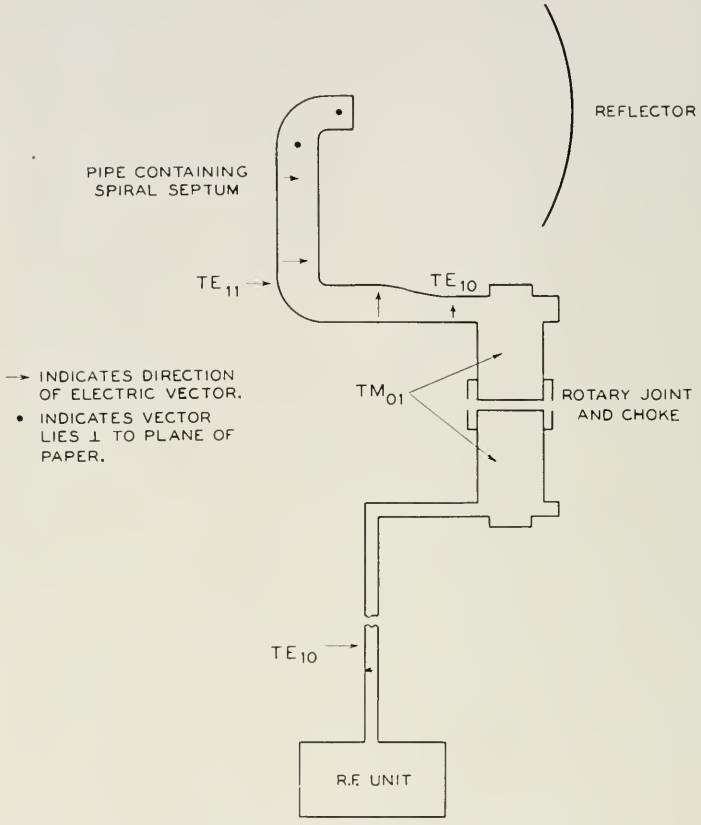


Fig. 43—SL Radar Antenna—Wave Guide Path.

The energy passes the rotary joint in this mode; choke labyrinths are provided at the joint to minimize radio frequency leakage. The energy then flows through another transducer, from TM_{01} mode back to TE_{10} mode. The lower horizontal portion of the feed pipe immediately tapers to round guide, the mode being now TE_{11}. Next the energy transverses a 90° elbow, which is a standard 90° pipe casting, and enters the vertical section im-

mediately below the feed aperture. The E vector is in the plane of the paper at this point. However, the ensuing vertical section is fitted with a spiral septum. This gradually rotates the plane of polarization until at the top of this pipe the E vector is perpendicular to the plane of the paper. Thus, after transversing another 90° pipe bend, the energy emerges horizontally polarized, to feed the main reflector.

Specific electrical characteristics of the SL antenna are:

Polarization—Horizontal
Horizontal Half Power Beamwidth—6°
Vertical Half Power Beamwidth—12°
Gain—about 22 db.

14.3 *The SJ Submarine Radar Antenna*

It had long been expected that one of the early offensive weapons of the war would be the submarine. It was therefore natural that early in the history of radar the need for practical submarine radars was felt. The principal components of this need were twofold, to provide warning of approaching enemies and to obtain torpedo fire control data. The SJ Submarine Radar was the first to be designed principally for the torpedo fire control function.

Work on the SJ system was under way considerably before Pearl Harbor. When this work was initiated the advantages of lobing fire control systems were clearly recognized, but no lobing antennas appropriate for submarine use had been developed. Requirements on such an antenna were obviously severe, for in addition to fulfilling fairly stringent electrical conditions, it would have to withstand very large forces due to water resistance and pressure.

The difficulties evident at the outset of the work were overcome by an ingenious adaptation of the simple waveguide feed. It was recognized that a shift of the feed in the focal plane of a reflector would cause a beam shift. Why not, then, use two waveguide feeds side by side to produce the two nearly coincident beams required in a lobing antenna? When this was tried it was found to work as expected.

It remained to devise a means of switching from one waveguide feed to the other with the desired rapidity. This in itself was no simple problem, but was solved by applying principles learned through work on waveguide filters. The switch at first employed was essentially a branching filter at the junction of the single antenna feed line and the line to each feed aperture. Both branches of this filter were carefully tuned to the same frequency, that of the radar. The switching was performed by the insertion of small rapidly rotating pins successively into the resonant cavities of the

two filters (Fig. 44). Presence of the pins in one of the filters detuned it and therefore prevented power from flowing through it. Rotation of the pins accordingly produced switching as desired.

In a later modification of this switch the same general principles were used but resonant irises rather than resonant cavities were employed.

The SJ Submarine Radar was in use at a comparatively early date in the war and saw much service with the Pacific submarine fleet. Despite some early doubts, submarine commanders were soon convinced of its powers.

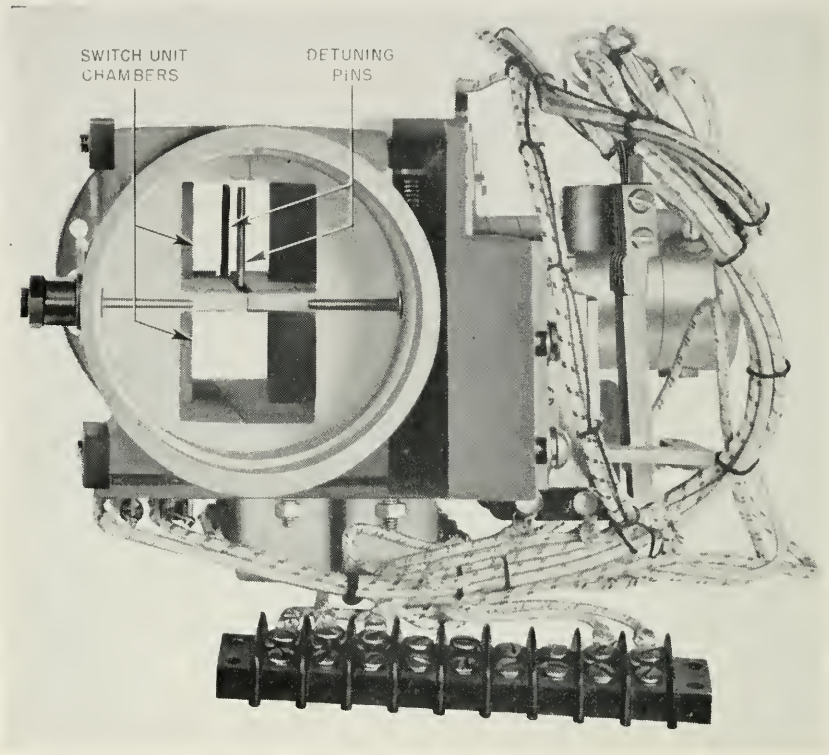


Fig. 44—The SJ Tuned Cavity Switch.

It is believed that in the majority of cases it replaced the periscope as the principle fire control instrument. In addition it served as a valuable and unprecedented aid to navigation.

It is interesting and relevant to quote from two letters to Laboratories engineers concerning the SJ. One dated October 3, 1943, from the radar officer of a submarine stated that there were twenty "setting sun" flags painted on the conning tower and asked the engineer to "let your mind dwell on the fact that you helped to put more than 50% of those flags there".

The commander of another submarine wrote in a similar vein, "You can rest assured that we don't regard your gear as a bushy-brain space taker, but a very essential part of our armament".

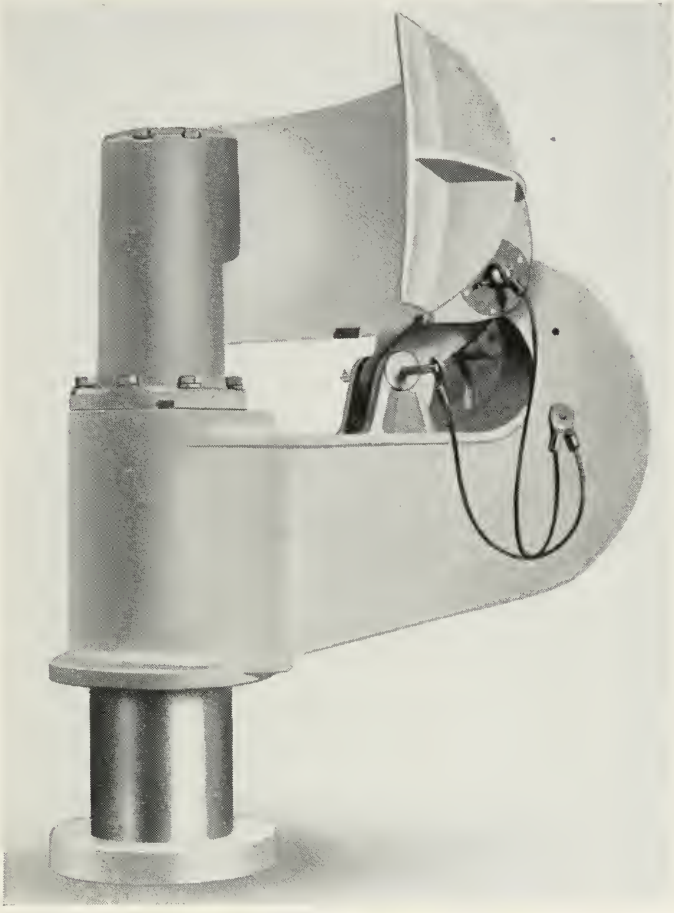


Fig. 45—The SJ Submarine Radar Antenna.

Figure 45 is a photograph of an SJ antenna. Its principal electrical characteristics are as follows:

Gain > 19 db
Horizontal Half Power Beamwidth 8°
Vertical Half Power Beamwidth 18°
Vertical Beam Character—Some upward radiation
Lobe Switching Beam Separation—approximately 5°
Gain reduction at beam cross-over < 1 db
Polarization—Horizontal

14.4 *The Modified SJ/Mark 27 Radar Antenna*

The SJ antenna described above performed a remarkable and timely fire control job as a lobing antenna but was found to be unsatisfactory when rotated continuously to produce a Plan Position Indicator (PPI) presentation. In the PPI method of presentation range and angle are presented as radius and angle on the oscilloscope screen. Consequently a realistic map of the strategic situation is produced. This map is easily spoiled by false signals due to large minor lobes of the antenna.

Since it had been established that the PPI picture was valuable for navigation and warning as well as for target selection it was decided to modify the antenna in a way that would reduce these undesirably high minor lobes. These were evidently due principally to the shadowing effect of the massively built double primary feed. Accordingly a new reflector was designed which in combination with a slightly modified feed provided a much improved pattern.

The new reflector was different in configuration principally in that it was a partially offset section of a paraboloid. The reflector surface was also markedly different in character since it was built as a grating rather than a solid surface. This reduced water drag on the antenna. In addition the grating was less visible at a distance, an advantage that is obviously appreciable when the antenna is the only object above the water.

This modified antenna was used not only on submarines as part of the SJ-1 radar but also on surface vessels as the Mark 27 Radar Antenna. Figure 46 shows one of these antennas. Its electrical characteristics are as follows:

Gain $>$ 20 db
Horizontal Half Power Beamwidth = 8°
Vertical Half Power Beamwidth = 17°
Vertical Beam Character—Some upward radiation
Lobe Switching Beam Separation—approximately 5°
Gain reduction at beam cross-over $<$ 1 db
Polarization—Horizontal

14.5 *The SH and Mark 16 Antenna*[12]

The antennas designed for the SH and Mark 16 Radar Equipments are practically identical. The SH system was a shipborne combined fire control and search system, and the Mark 16 its land based counterpart was used by the Marine Corps for directing shore batteries.

These systems operated at 9.8 cm. The requirement that the system, operate as a fire control as well as a search system imposed some rather stringent mechanical requirements on the antenna. For search purposes, the antenna was rotated at 180 rpm, and indications were presented on a plan position indicator. For fire control data, slow, accurately controlled motion was required. Bearing accuracy is attained by lobe switching in

[12] Written by R. J. Philipps.

much the same manner as in the SJ and SJ-1 antennas previously described.

The antenna is illustrated in Fig. 47. With the SH system, the unit is mast mounted; for the Mark 16, the unit is mounted atop a 50 foot steel

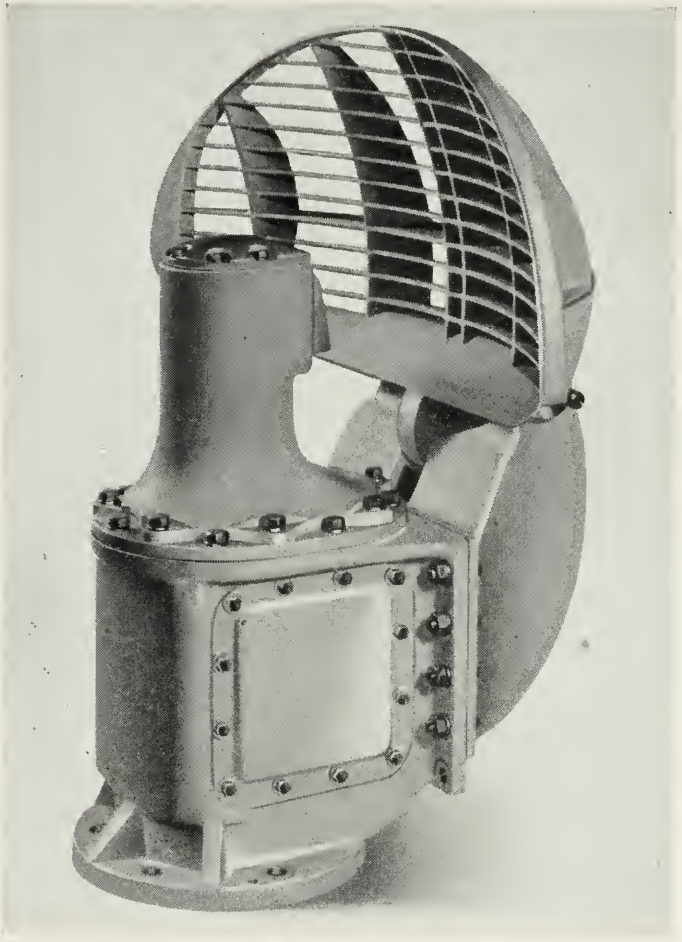


Fig. 46—The SJ-1/Mark 27 Submarine Radar Antenna.

tower which can be erected in a few hours with a minimum of personnel. The electrical characteristics are as follows:

Gain—21. db
Reflector Dimensions 30″ W x 20″ H
Horizontal beam width—7.5°
Vertical beam width—12°
Lobe separation—5° approximately
Loss in gain at lobe crossover—1 db approximately
Scan—(1) 360°, at 180 rpm for PPI operation
(2) 360°, at approximately 1 rpm for accurate azimuth readings, with lobe switching

SH systems were most successfully used in invasion operations in the Aleutians. They were installed on landing craft, and the use of the high

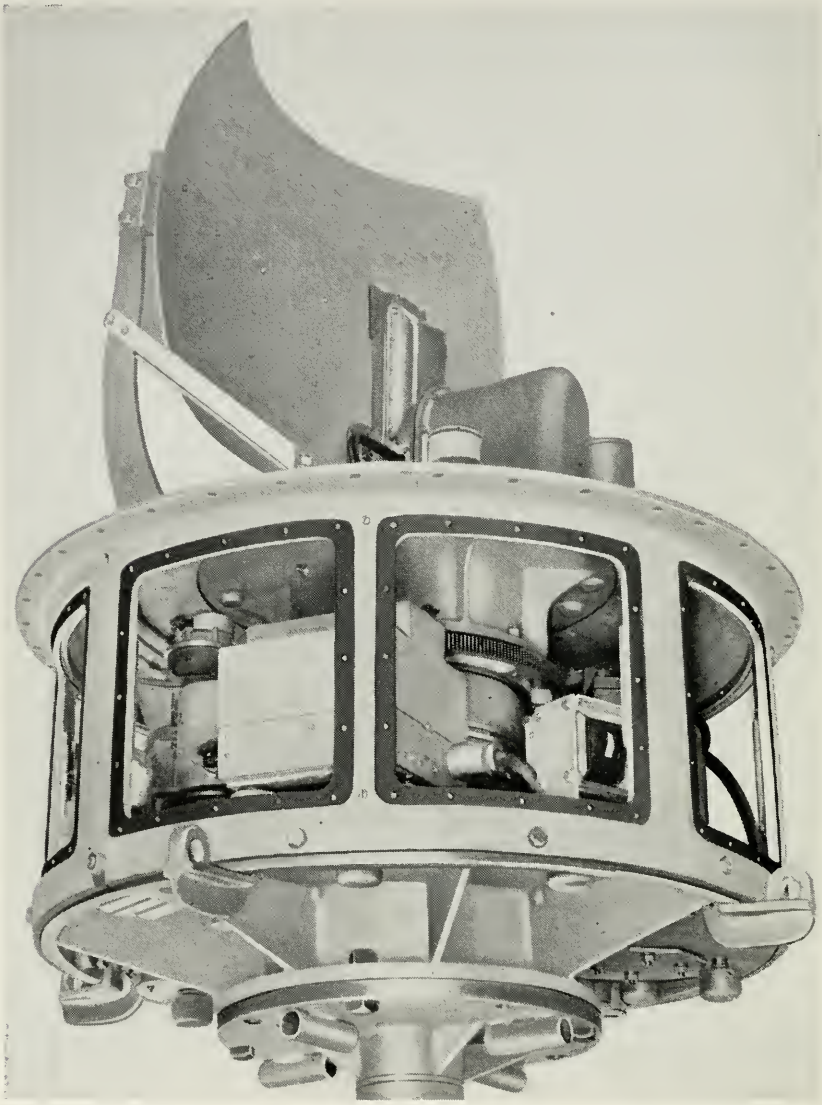


Fig. 47—SH Antenna.

speed scan enabled the craft to check constantly their relative positions in the dense fogs encountered during the landing operations.

14.6 *Antennas for Early Fire Control Radars*[13]

The first radars to be produced in quantity for fire control on naval vessels were the Mark 1, Mark 3 and Mark 4 (originally designated FA, FC and FD). These radars were used to obtain the position of the target with sufficient accuracy to permit computation of the firing data required by the guns. The first two (Mark 1 and Mark 3) were used against enemy surface targets while the Mark 4 Radar was a dual purpose system for use against both surface and aircraft targets. These radars were described in detail in an earlier issue.[14] However, photographs of the antennas and pertinent information on the antenna characteristics are repeated herein for the sake of completeness. (See Table B and Figures 48, 49 and 50.)

TABLE B

	Radar			
	Mark 1	Mark 3		Mark 4
Dimensions	6′ x 6′	3′ x 12′	6′ x 6′	6′ x 7′
Operating Frequency	500 or 700 MC	680–720 MC		680–720 MC
Beam Width in Degrees (Between half power points one way.)				
Azimuth	12°	6°	12°	12°
Elevation	14°	30°	14°	12°
Antenna Gain	22 db	22 db	22 db	22.5 db.
Beam Shift in Degrees				
Azimuth	0°	±1.5°	±3°	±3°
Elevation	0°	0°	0°	±3°

An antenna quite similar to the Mark 3, 6 ft. x 6 ft. antenna, was also used on Radio Set SCR-296 for the Army. This equipment was similar to the Mark 3 in operating characteristics but was designed mechanically for fixed installations at shore points for the direction of coast artillery gun fire. For these installations the antenna was mounted on an amplidyne controlled turntable located on a high steel tower. The entire antenna and turntable was housed within a cylindrical wooden structure resembling a water tower. Equipments of this type were used as a part of the coastal defense system of the United States, Hawaiian Islands, Aleutian Islands and Panama.

[13] Written by W. H. C. Higgins.

[14] "Early Fire Control Radars for Naval Vessels," W. C. Tinus and W. H. C. Higgins, B. S. T. J.

14.7 *A Shipborne Anti-Aircraft Fire Control Antenna*[15]

A Shipborne Anti-Aircraft Fire Control Antenna is shown in Fig. 51. This antenna consists of two main horizontal cylindrical parabolas in each

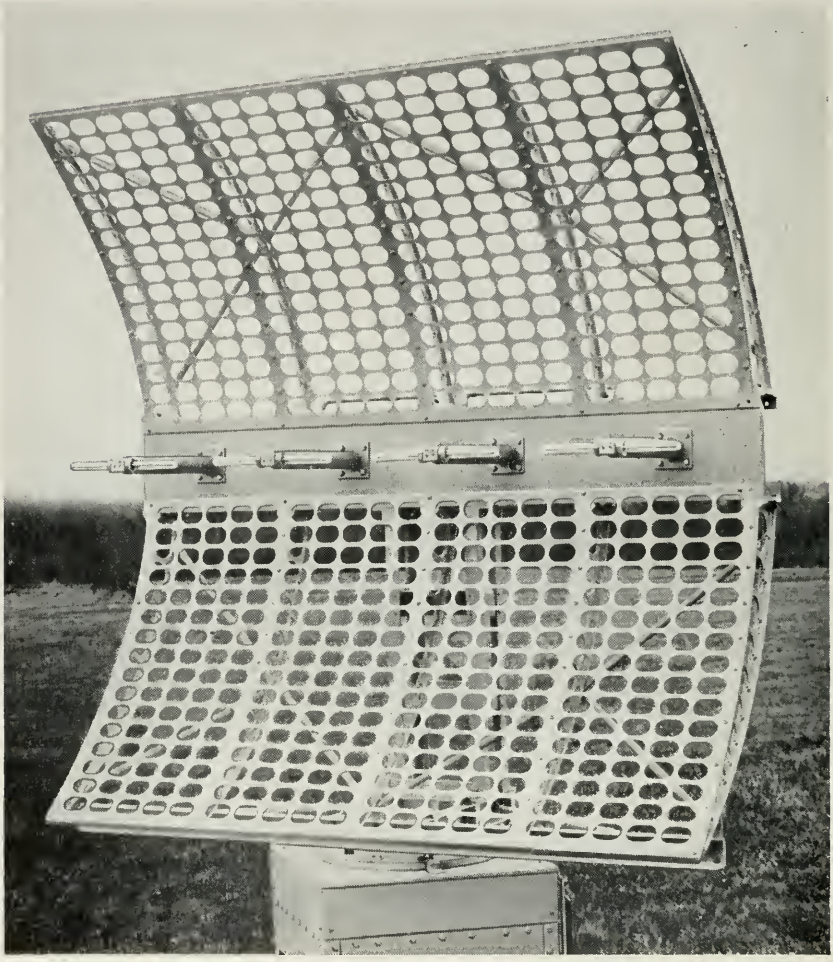


Fig. 48—Mark 1 Antenna.

of which two groups of four half-wave dipoles are mounted with their axes in a horizontal line at the focus of the parabolic reflectors. The four groups of dipoles are connected by coaxial lines on the back of the antenna to a lobe

[15] Written by C. A. Warren.

switcher, which is a motor driven capacitor that has a single rotor plate and four stator plates, one for each group of dipoles. The phase shift introduced into the four feed lines by the lobe switching mechanism causes the antenna beam to be "lobed" or successively shifted to the right, up, left and down as the rotor of the capacitor turns through 360 degrees.

Mounted centrally on the front of the antenna at the junction of the two parabolic antennas is a smaller auxiliary antenna consisting of two dipole elements and a parabolic reflector, the purpose of which is to reduce the minor lobes that are present in the main antenna beam. The auxiliary

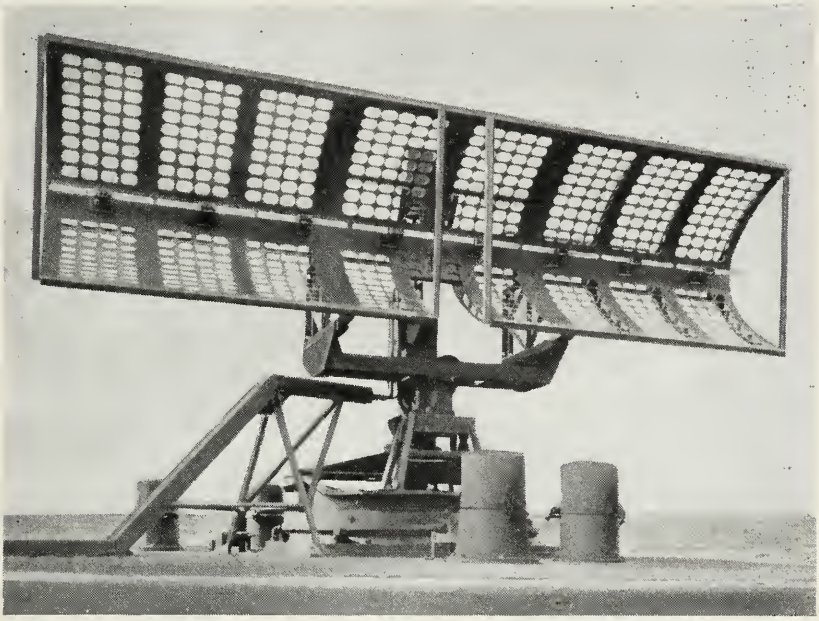


Fig. 49—Mark 3 Radar Antenna on Battleship New Jersey.

antenna beam is not lobe switched and is sufficiently broad in both the horizontal and vertical planes to overlap both the main antenna beam and the first minor lobes. The auxiliary antenna feed is so designed that its field is in phase with the field of the main beam of the main antenna. This causes the feed of the auxiliary antenna to "add" to the field of the main antenna in the region of its main beam, but to subtract from the field in the region of its first minor lobes. This occurs because the phase of the first minor lobes differs by 180 degrees from that of the main beam. As a result, the field of the main beam is increased and the first minor lobes are greatly

reduced. By reducing these minor lobes to a low value, the region around the main beam is free of lobes, thus greatly reducing the possibility of false tracking due to "cross overs" between the main beam and the minor lobes.

14.8 *The Polyrod Fire Control Antenna*

The Polyrod Fire Control antenna is an array scanner employing essentially the same principles as those used in the multiple unit steerable antenna

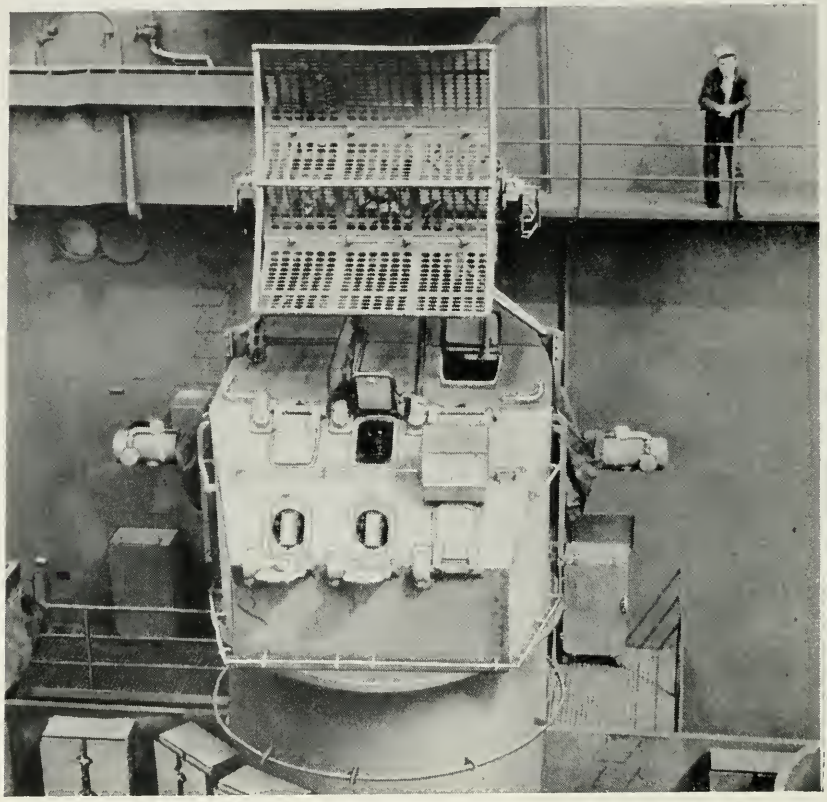


Fig. 50—Mark 4 Radar Antenna on Battleship Tennessee.

system (MUSA) developed before the war for short-wave transatlantic telephony. Some of these principles have been discussed in Sec. 12.2. That they could be applied with such success in the microwave region was due to a firm grounding in waveguide techniques, to the invention of the polyrod antenna and the rotary phase changer, and especially to excellent technical work on the part of research, development and production personnel. It is perhaps one of the most remarkable achievements of wartime

radar that the polyrod antenna emerged to fill the rapid scanning need a early and as well developed as it did.

The Polyrod Fire Control antenna is a horizontal array of fourteen identical fixed elements, each element being a vertical array of three polyrods. Energy is distributed to the elements through a waveguide manifold. The phase of each element is controlled and changed to produce the desired scan by means of thirteen rotary phase changers. These phase shifters are

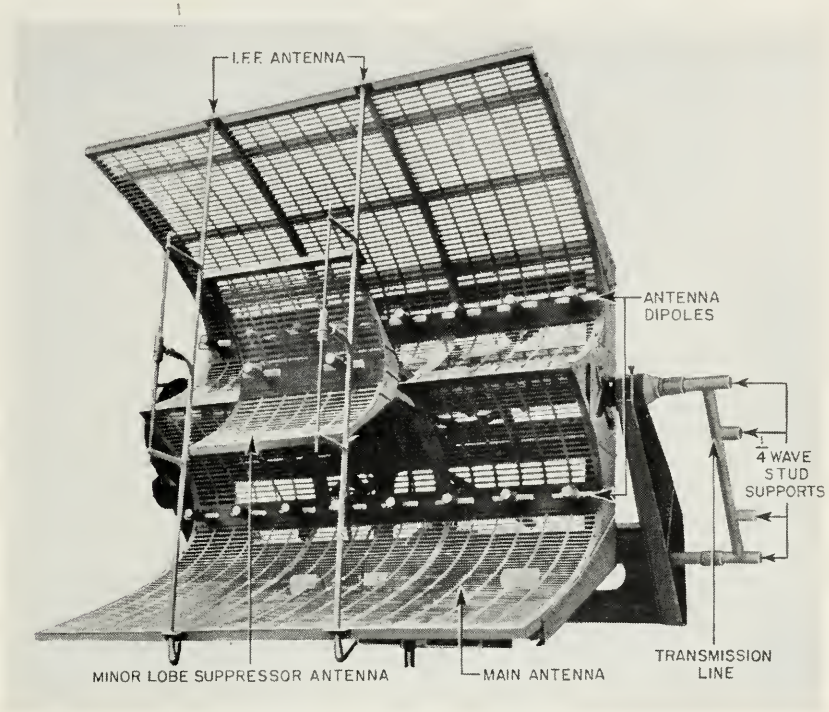


Fig. 51.—Shipborne Anti-Aircraft Fire Control Antenna

geared together and driven synchronously. Figure 52 is a schematic diagram of the waveguide and phase changer circuits.

Figure 39 shows an experimental polyrod antenna under test at Holmdel. Figure 53 is another view of the Polyrod antenna.

14.9 *The Rocking Horse Fire Control Antenna*

It was long recognized that an important direction of Radar development lay towards shorter waves. This is particularly true for fire control antennas where narrow, easily controlled beams rather than great ranges are needed. The Polyrod antenna had pretty thoroughly demon-

strated the value of rapid scanning, yet the problem of producing a rapid scanning higher frequency antenna of nearly equal dimensions was a new and different one.

Several possible solutions to this problem were known. The array technique applied so effectively to the polyrod antenna could have been applied here also, but only at the expense of many more elements and greater complexity.

After much preliminary work it was finally concluded that a mechanically scanning antenna, the "rocking horse," provided the best solution to the higher frequency scanning problem. This solution is practical and relatively simple.

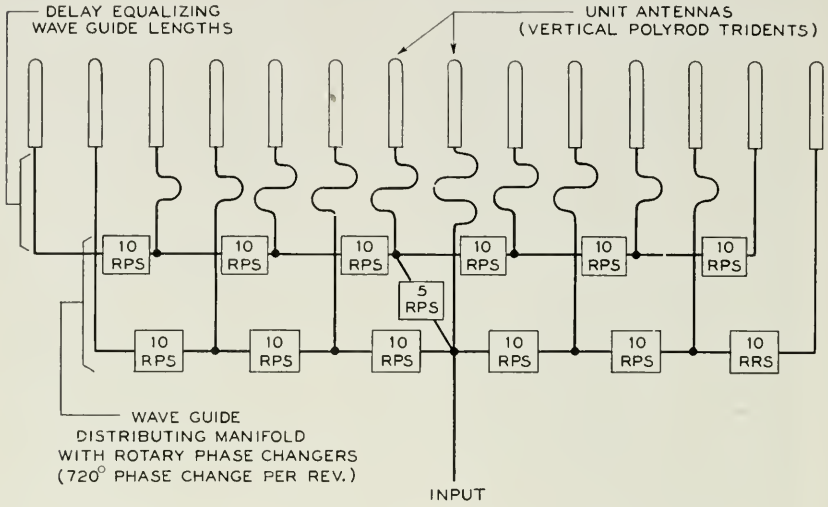


Fig. 52.—Schematic Diagram of Polyrod Fire Control Antenna.

The operation of the rocking horse is described in Sec. 12.1. It is essentially a carefully designed and firmly built paraboloidal antenna which oscillates rapidly through the scanning sector. Its oscillation is dynamically balanced to eliminate undesirable vibration.

Figure 54 is a photograph of a production model of the rocking horse antenna.

14.10 *The Mark 19 Radar Antenna*[16]

In Anti-aircraft Fire Control Radar Systems for Heavy Machine Guns it is necessary to employ a highly directive antenna and to obtain continuous rapid comparison of the received signals on a number of beam positions

[16] Sections 14.10, 14.11 and 14.12 were written by F. E. Nimmcke.

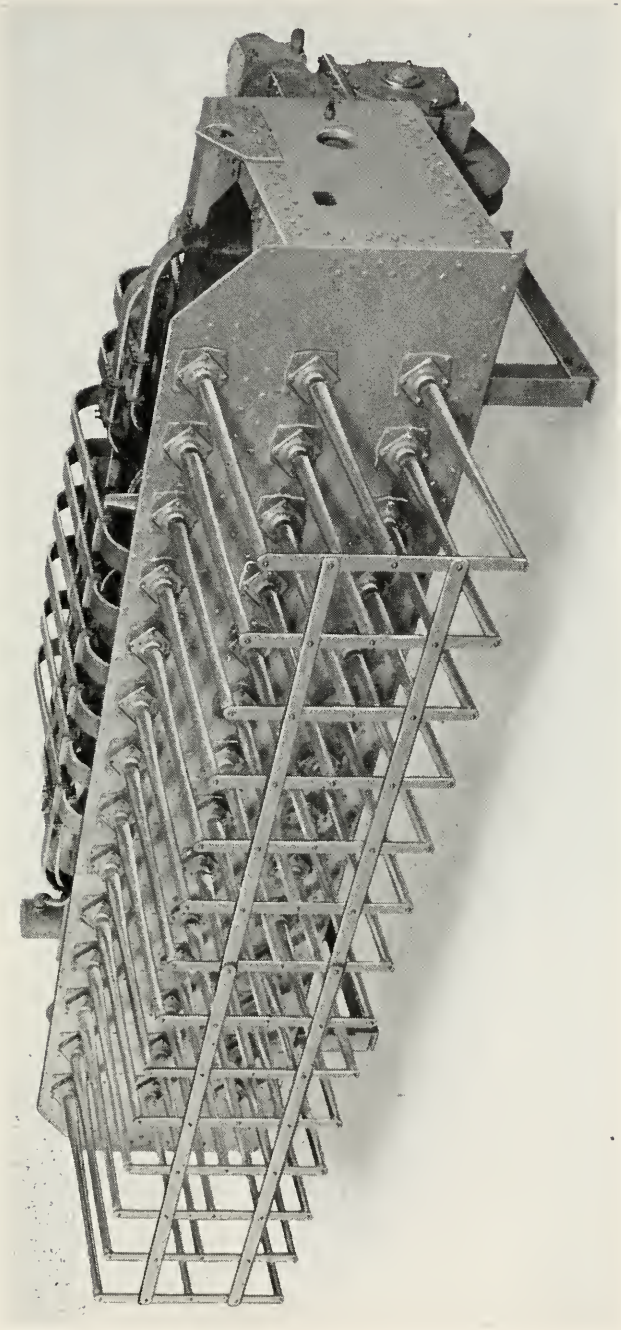


Fig. 53.—Polyrod Fire Control Antenna.

as discussed in Section 11.2. Such an antenna is also required to obtain the high angular precision for anti-aircraft fire control. These requirements are achieved by the use of a conical scanning system. The beam from the antenna describes a narrow cone and the deviation of the axis of the cone from the line of sight to the target can be determined and measured by the phase difference between the amplitude modulated received signal and the frequency of the reference generator associated with the

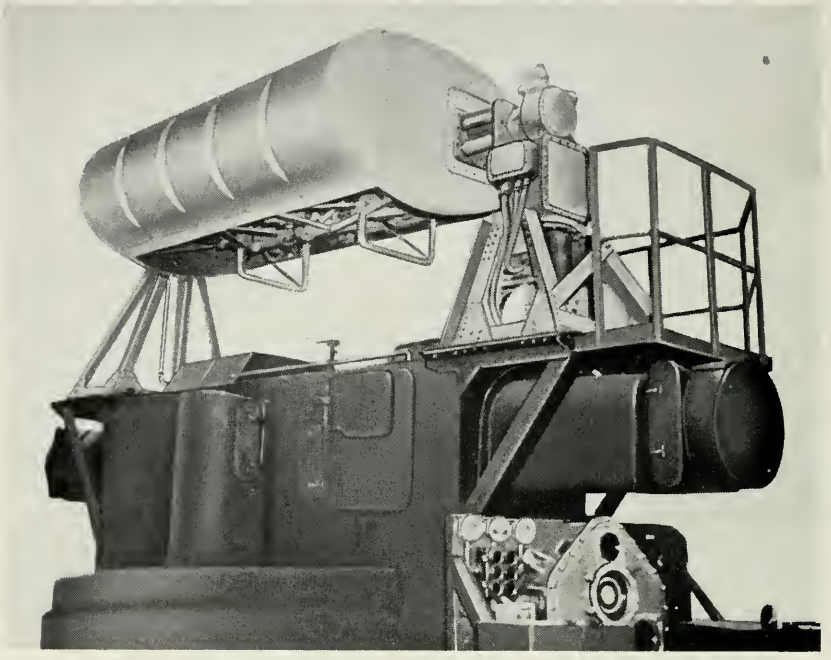


Fig. 54.—Rocking Horse Fire Control Antenna.

antenna. This information is presented to the pointer-trainer at the director in the form of a wandering dot on an oscilloscope.

The antennas described in sections 14.10, 14.11 and 14.12 were all designed by the Bell Laboratories as anti-aircraft fire control radar systems, particularly for directing heavy machine guns. They were designed for use on all types of Naval surface warships.

In Radar Equipment Mark 19, the first system to be associated with the control of 1.1 inch and 40 mm anti-aircraft machine guns, the antenna was designed for operation in the 10 cm region. This antenna consisted of a spinning half dipole with a coaxial transmission line feed. The antenna

was driven by 115-volt, 60 cycle, single phase motor to which was coupled a two-phase reference voltage generator. The motor rotated at approximately 1800 rpm which resulted in a scanning rate of 30 cycles per second. This antenna was used with a 24-inch spun steel parabolic reflector which provided, at the 3 db point, a beam width of approximately 11° and a beam shift of 8.5° making a total beam width of approximately 20° when scanning. The minor lobes were down more than 17 db (one way) from the maximum; and the gain of this antenna was 21 db. This antenna assembly

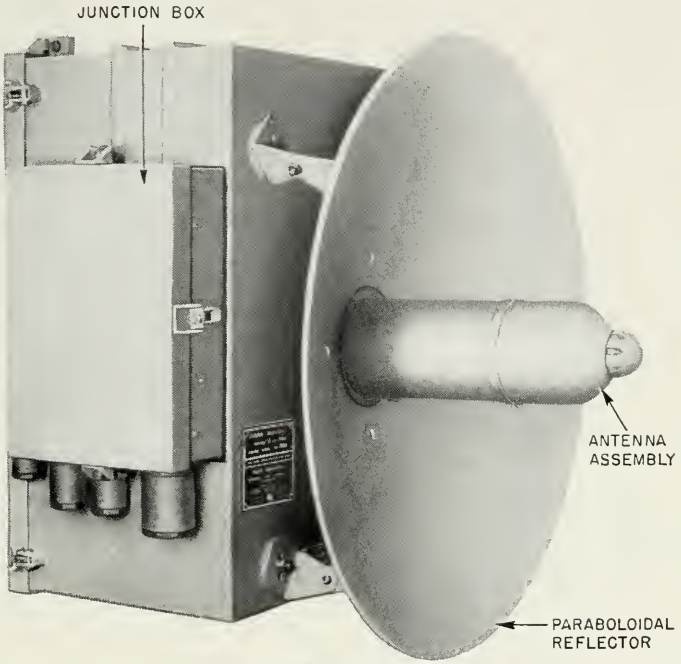


Fig. 55—Mark 19 Antenna.

was integral with a transmitter-receiver (Fig. 55) which was mounted on the associated gun director. Consequently, the size of the reflector was limited by requirements for unobstructed vision for the operators in the director. As a matter of fact, for this type of radar system serious consideration must be given to the size and weight of the antenna and associated components.

14.11 *The Mark 28 Radar Antenna*

The beam from the antenna used in Radar Equipment Mark 19 was relatively broad and to improve target resolution, the diameter of the

reflector for the antenna in Mark 28 was approximately doubled. The Mark 28 is a 10 cm system and employs a conical scanning antenna similar to that described for Mark 19. The essential difference is that the spun steel parabolic reflector is 45 inches in diameter which provides a beam width of approximately 6.5° and a beam shift of 4.5° making a total of 11°.

Fig. 56—Mark 28 Antenna Mounted on 40 MM Gun.

The minor lobes are down more than 17 db (one way) from the maximum; and the gain of this antenna is 26 db. It was found necessary to perforate the reflector of this dimension in order to reduce deflection caused by gun blast and by wind drag on the antenna assembly. The antenna assembly for Radar Equipment Mark 28 is shown in Fig. 56. This assembly is shown mounted on a 40 mm Gun.

14.12 *A 3 CM Anti-Aircraft Radar Antenna.*

To obtain greater discrimination between a given target and other targets, or between a target and its surroundings, the wavelength was reduced to the 3 cm region. An antenna for this wavelength was designed to employ the conical scan principle. In this case the parabolic reflector was 30 inches in diameter and transmitted a beam approximately 3° wide at the 3db point with a beam shift of 1.5° making a total of 4.5° with the antenna scanning. The minor lobes are down more than 22 db (one way) from the maximum; and the gain of this antenna is 35 db.

In the 3 cm system in which a Cutler feed was used, the axis of the beam was rotated in an orbit by "nutation" about the mechanical axis of the antenna. This was accomplished by passing circular waveguide through the hollow shaft of the driving motor. The rear end of the feed (choke coupling end) was fixed in a ball pivot while the center (near the reflector) was off set the proper amount to develop the required beam shift. This off set was produced by a rotating eccentric driven by the motor. The latter was a 440 volt, 60 cycle, 3 phase motor rotating at approximately 1800 rpm which resulted in a scanning rate of 30 cycles per second. The two-phase reference voltage generator was integral with the driving motor.

It was found necessary at these radio frequencies to use a cast aluminum reflector and to machine the reflecting surface to close tolerances in order to attain the consistency in beam width and beam direction required for accurate pointing. An antenna assembly for the 3 cm anti-aircraft radar is shown in Fig. 57.

15. Land Based Radar Antennas

15.1 *The SCR-545 Radar "Search" and "Track" Antennas*[17]

The SCR-545 Radar Set was developed at the Army's request to meet the urgent need for a radar set to detect aircraft and provide accurate target tracking data for the direction of anti-aircraft guns.

This use required that a narrow beam tracking antenna be employed to achieve the necessary tracking accuracy, furthermore, a narrow beam antenna suitable for accurate tracking has a very limited field of view and requires additional facilities for target acquisition. This was provided by the search antenna which has a relatively large field of view and is provided with facilities for centering the target in its field of view. These two antennas are integrated into a single mechanical structure and both radar axes coincide.

The "Search" antenna operates in the 200 mc band and is com-

[17] Section 15.1 was written by A. L. Robinson.

posed of an array of 16 quarter wave dipoles spaced 0.1 wave-length in front of a flat metal reflector. All feed system lines and impedance matching devices are made up of coaxial transmission line sections. The array is divided into four quarters, each being fed from the lobe switching mechanism. This division is required to permit lobe switching in both horizontal and vertical planes. The function of the lobe switching mecha-

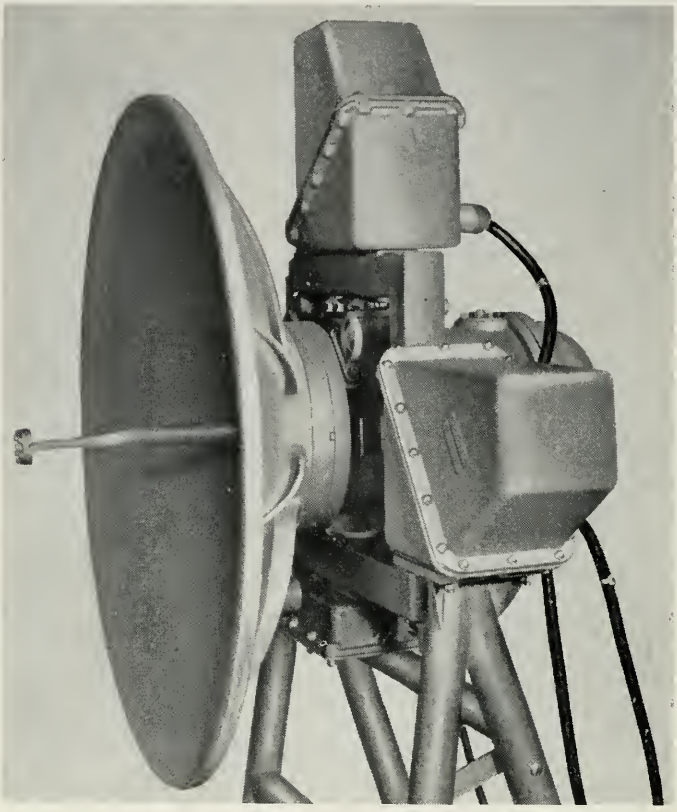


Fig. 57.—3CM Anti-Aircraft Radar Antenna.

nism is to introduce a particular phase shift in the excitation of the elements of one half of the antenna with respect to the other half. The theory of this type of lobe switching is discussed in section 11.1. The antenna beam spends approximately one quarter of a lobing cycle in each one of the four lobe positions. Each of the four lobe positions has the same radiated field intensity along the antenna axis and therefore when a target is on axis equal signals will be received from all four lobe positions.

The "Track" antenna operates in the 10 cm. region and consists of a reflector which is a parabola or revolution, 57 inches in diameter, illuminated by a source of energy emerging from a round waveguide in the lobing mechanism. Conical lobing is achieved by rotating the source of energy around the parabola axis in the focal plane of the parabola. Conical lobing is discussed in section 11.2. The round waveguide forming the source is filled with a specially shaped polystyrene core to control the illumination of the parabola and to seal the feed system against the weather. The radio frequency power is fed through coaxial transmission line to a coaxial-waveguide transition which is attached to the lobing mechanism.

The "Search" and "Track" antenna lobing mechanisms are synchronized and driven by a common motor.

The radio frequency power for both antennas is transmitted through a single specially constructed coaxial transmission line to the common antenna structure, where a coaxial transmission line filter separates the power for each antenna.

Figure 58 is a photograph of a production model of the SCR-545 Radar Set. The principal electrical characteristics of the antennas are tabulated below:

	Antennas	
	Search	Track
Gain	14.5 db	30 db
Horizontal Beamwidth	23.5°	5°
Vertical Beamwidth	25.5°	5°
Polarization	Horizontal	Vertical
Type of Lobing	Lobe switching	Conical lobing
Angle between lobe positions	10°	3°
Lobing rate	60 cycles/sec.	60 cycles/sec.

The SCR-545 played an important part in the Italian campaign, particularly in helping to secure the Anzio Beach Head area, as well as combating the "V" bombs in Belgium. However the majority of SCR-545 equipments were sent to the Pacific Theater of Operations and played an important part in operations on Leyte, Saipan, Iwo Jima, and Okinawa.

15.2 *The AN/TPS-1A Portable Search Antenna*[18]

In order to provide early warning information for advanced units, a light weight, readily transportable radar was designed under Signal Corps contract.

[18] Written by R. E. Crane.

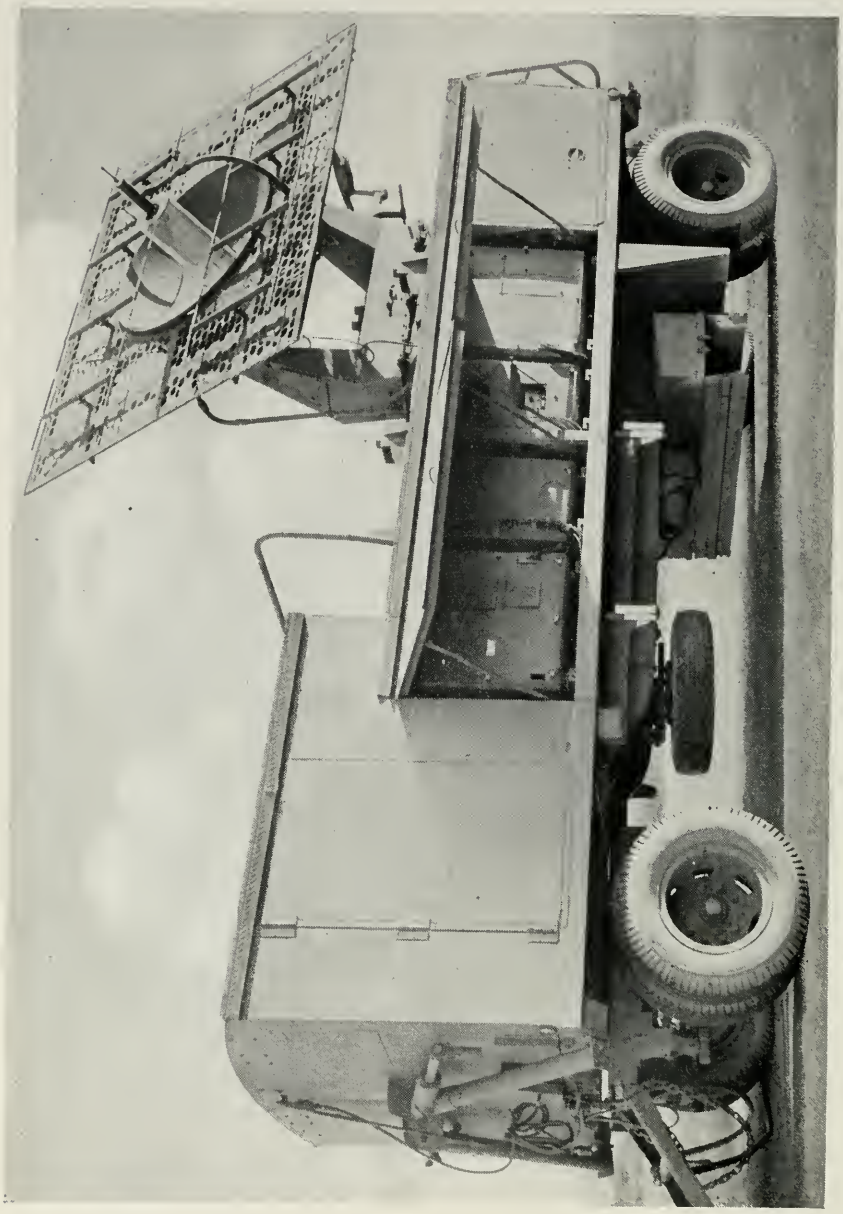


Fig. 58—SCR/545 Antenna.

The objective was to obtain as long range early warning as possible with moderate accurracy of location. Emphasis was placed on detection of low flying planes.

The objectives for the set indicated that the antenna should be built as large as reasonable and placed as high as reasonable for a portable set. Some latitude in choice of frequency was permitted at first. For ruggedness and reliability reasons which seemed controlling at the time, the frequency was pushed as high as possible with vacuum tube detectors and R.F. amplifiers. This was finally set at 1080 mc.

Fig. 59—AN/TPS-1A Antenna.

The antenna as finally produced was 15 ft. in width and 4 ft. in height. The reflecting surface was paraboloidal. The mouth of the feed horn was approximately at the focus of the generating parabola. The feedhorn was excited by a probe consisting of the inner conductor of the coaxial transmission line extended through the side of the horn and suitably shaped. To reduce side lobes and back radiation the feedhorn was dimensioned to taper the illumination so that it was reduced about 10 db in the horizontal and vertical planes at the edges of the reflector. Dimensions of probe and exact location of feed, etc. were determined empirically to secure acceptable impedance over the frequency band needed. This band, covered by spot frequency magnetrons, was approximately $\pm 2.5\%$ from mid frequency.

Figure 59 shows the antenna in place on top of the set.

The characteristics of this antenna are summarized below:

Gain	27.3 db.
Horizontal Half Power Beamwidth	4.4°
Vertical Half Power Beamwidth	12.6°
Vertical Beam Characteristic	Symmetrical
Polarization	Horizontal
Impedance (SWR over ±2.5% band)	<4.0 db

16. Airborne Radar Antennas

16.1 *The AN/APS-4 Antenna*[19]

AN/APS-4 was designed to provide the Navy's carrier-based planes with a high performance high resolution radar for search against surface and airborne targets, navigation and interception of enemy planes under conditions of fog and darkness. For this service, weight was an all important consideration and throughout a production schedule that by V-J day was approaching 15,000 units, changes to reduce weight were constantly being introduced. In late production the antenna was responsible for 19 lbs. out of a total equipment weight of 164 lbs. The military requirements called for a scan covering 150° in azimuth ahead of the plane and 30° above and below the horizontal plane in elevation. To meet this requirement a Cutler feed and a parabolic reflector of 6.3″ focal length and 14½″ diameter was selected. Scanning in azimuth was performed by oscillating reflector and feed through the required 150° while elevation scan was performed by tilting the reflector. Beam pattern was good for all tilt angles. In early flight tests the altitude line on the B scope due to reflection from the sea beneath was found to be a serious detriment to the performance of the set. To reduce this, a feed with elongated slots designed for an elliptical reflector was tried and found to give an improvement even when used with the approximately round reflector. The elliptical reflector was also tried, but did not improve the performance sufficiently to justify the increased size.

As will be noted in Fig. 60, the course of the mechanical development brought the horizontal pivot of the reflector to the form of small ears projecting through the parabola. No appreciable deterioration of the beam pattern due to this unorthodox expedient was noted.

The equipment as a whole was built into a bomb-shaped container hung in the bomb rack on the underside of the wing. Various accidents resulted in this container being torn off the wing in a crash landing in water or dropped on the deck of the carrier. After these mishaps, the equipment was frequently found to be in good working order with little or no repair required.

[19] Written by F. C. Willis.

Gain	28 db
Beamwidth	6° approx. circular
Polarization	Horizontal
Scan	Mechanical
Scanning Sector—Azimuth	150°
Scanning Sector—Elevation	60°
Scanning Rate	one per sec.
Total weight	19 lbs.

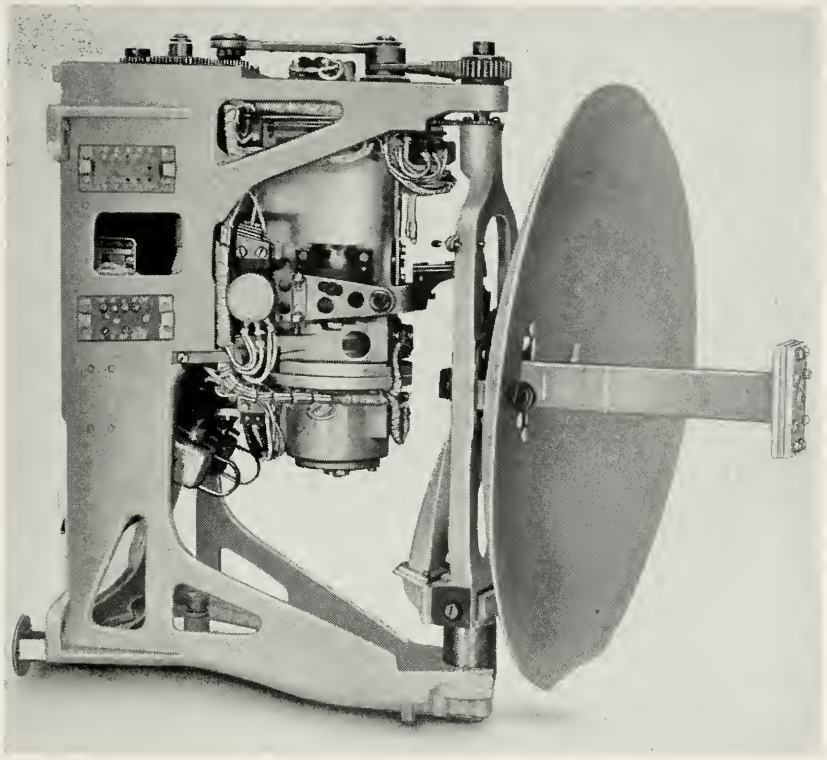


Fig. 60—AN/APS-4 Antenna.

16.2 *The SCR-520, SCR-717 and SCR-720 Antennas*[20]

The antenna shown in Fig. 61 is typical of the type used with the SCR-520 and SCR-720 aircraft interception (night fighter) airborne radar equipment, as well as the SCR-717 sea search and anti-submarine airborne radar equipment. The parabolic reflector is 29 inches in diameter and produces a radiation beam about 10° wide. The absolute gain is approximately 25 db. RF energy is supplied to a pressurized emitter through a pressurized transmission line system which includes a rotary joint located on the ver-

[20] Written by J. F. Morrison.

tical axis and a tilt joint on the horizontal axis. Either vertical or horizontal polarization can be used by rotating the mounting position of the emitter. Vertical polarization is preferred for aircraft interception work and horizontal polarization is preferred for sea search work.

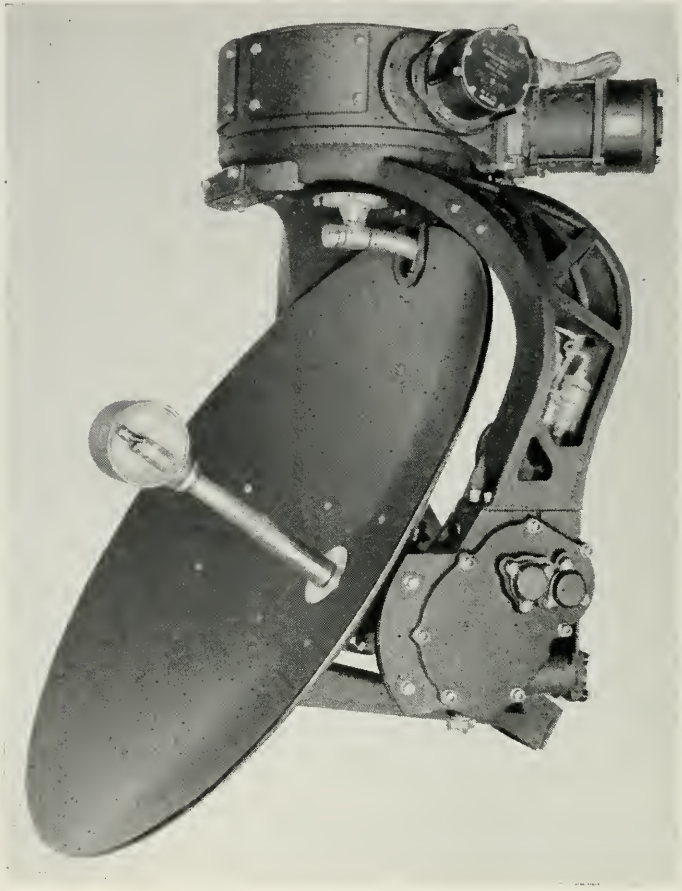


Fig. 61—SCR-520 Antenna.

For aircraft interception the military services desired to scan rapidly a large solid angle forward of the pursuing airplane, i.e. 90° right and left, 15° below and 50° above the line of flight. The data is presented to the operator in the form of both "B" and "C" presentations and for this purpose potentiometer data take-offs are provided on the antenna. The reflector is spun on a vertical axis at a rate of 360 rpm and at the same time it is

made to nod up and down about its horizontal axis by controllable amounts up to a total of 65° and at a rate of 30° per second.

In the sea search SCR-717 equipment, selsyn azimuth position data take-offs are provided which drive a PPI type of indicator presentation. The rotational speed about the vertical axis in this case is either 8 or 20 rpm as selected by the operator. The reflector can also be tilted about its horizontal axis above or below the line of flight as desired by the operator.

It will be noted that the emitter moves with the reflector and accordingly it is always located at the focal point throughout all orientations of the antenna.

16.3 *The AN/APQ-7 Radar Bombsight Antenna*[21]

Early experience in the use of bombing-through-overcast radar equipment indicated that a severe limitation in performance was to be expected as the result of the inadequate resolution offered by the then available airborne radar equipments. This lack of resolution accounted for gross errors in bombing where the target area was not ideal from a radar standpoint.

To meet this increased resolution requirement in range, the transmitted pulse width was shortened considerably. In attempting to increase the azimuthal resolution, higher frequencies of transmission were employed. This enabled an improvement in azimuthal resolution without resorting to larger radiating structures, a most important consideration on modern high speed military aircraft.

To extend the size of the radiating structure without penalizing the aircraft performance, the use of a linear scanning array which would exhibit high azimuthal resolution was considered. This array was originally conceived in a form suitable to mount within the existing aircraft wing and transmit through the leading edge. As development proceeded, the restrictions imposed on the antenna structure as well as the aircraft wing design resulted in the linear array scanner being housed in an appropriate separate air foil and attached to the aircraft fuselage (Fig. 62).

The above study resulted in the development of the AN/APQ-7 radar equipment, operating at the X-band of frequencies.[22] This equipment provided facilities for radar navigation and bombing.

The AN/APQ-7 antenna consisted of an array of 250 dipole structures spaced at $\frac{1}{2}$ wavelength intervals and energized by means of coupling probes extending into a variable width waveguide. The vertical pattern was arranged to exhibit a modified $\csc^2$ distribution by means of accurately shaped "flaps" attached to the assembly.

[21] Written by L. W. Morrison.

[22] A large part of the antenna development was carried out at the M. I. T. Radiation Laboratory.

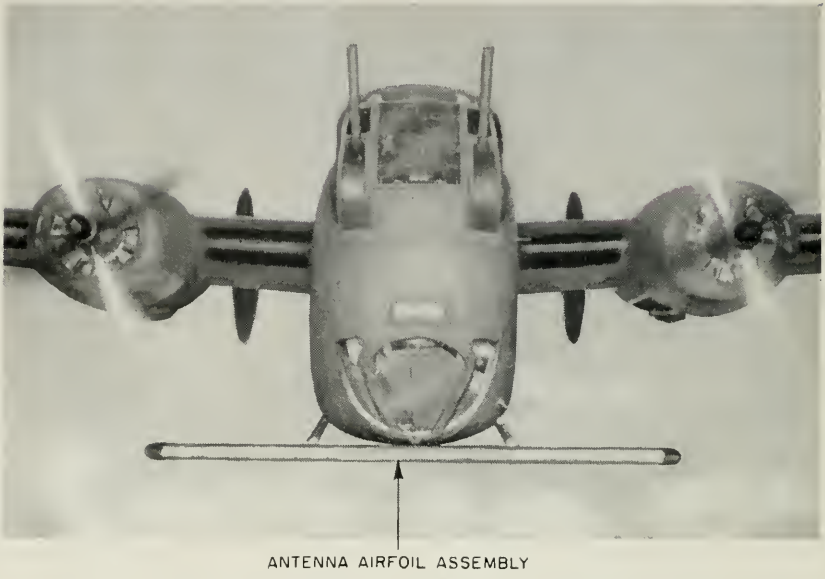


Fig. 62—AN/APQ-7 Antenna Mounted on B24 Bomber.

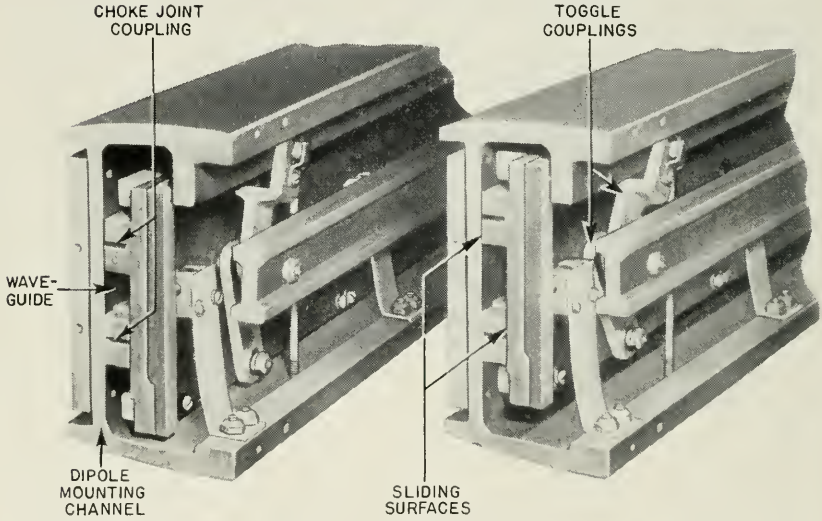


Fig. 63—AN/APQ-7 Antenna. Left—Contracted Wave Guide Assembly. Right—Expanded Wave Guide Assembly.

The scanning of the beam is accomplished by varying the width of the feed waveguide. This is accomplished by means of a motor driven actuated cam which drives a push rod extending along the waveguide assembly back

and forth. Toggle arms are attached to this push rod at frequently spaced intervals which provides the motion for varying the width of the waveguide while assuring precise parallelism of the side walls throughout its length (Fig. 63).

The normal range of horizontal scanning exhibited by this linear array, extends from a line perpendicular to the array to 30° in the direction of the feed. By alternately feeding each end, a total scanning range of ±30° from the perpendicular is achieved. Appropriate circuits to synchronize the indicator for this range are included.

The use of alternate end feed on the AN/APQ-7 antenna requires that the amount of energy fed to the individual dipoles is somewhat less than if a single end feed is employed.

The AN/APQ-7 antenna is $16\frac{1}{2}$ feet in length and weighs 180 pounds exclusive of air foil housing.

The following data applies:

Gain = 32.5 db
Horizontal beamwidth = 0.4°
Vertical beam characteristic = modified $\csc^2$
Scan—Array scanning
Scanning Sector—± 30° Horizontal
Scanning Rate = 45°/second

Acknowledgments

Contributors to the research and development of the radar antennas described in this paper included not only the great number of people directly concerned with these antennas but also the many people engaged in general research and development of microwave components and measuring techniques. A complete list of credits, therefore, will not be attempted.

In addition to the few individuals mentioned in footnotes throughout the paper, the authors would like to pay special tribute to the following co-workers in the Radio Research Department: C. B. H. Feldman who with the assistance of D. H. Ring made an outstanding contribution in the development of the polyrod array antenna; W. A. Tyrrell for his work on lobe switches; A. G. Fox, waveguide phase changers; A. P. King, paraboloids and horn antennas; A. C. Beck, submarine antennas; G. E. Mueller, polyrods.

Probability Functions for the Modulus and Angle of the Normal Complex Variate

By RAY S. HOYT

This paper deals mainly with various 'distribution functions' and 'cumulative distribution functions' pertaining to the modulus and to the angle of the 'normal' complex variate, for the case where the mean value of this variate is zero. Also, for auxiliary uses chiefly, the distribution function pertaining to the reciprocal of the modulus is included. For all of these various probability functions the paper derives convenient general formulas, and for four of the functions it supplies comprehensive sets of curves; furthur, it gives a table of computed values of the cumulative distribution function for the modulus, serving to verify the values computed by a different method in an earlier paper by the same author.[1]

Introduction

IN THE solution of problems relating to alternating current networks and transmission systems by means of the usual complex quantity method, any deviation of any quantity from its reference value is naturally a complex quantity, in general. If, further, the deviation is of a random nature and hence is variable in a random sense, then it constitutes a 'complex random variable,' or a 'complex variate,' the word 'variate' here meaning the same as 'random variable' (or 'chance variable'—though, on the whole, 'random variable' seems preferable to 'chance variable' and is more widely used).

Although a complex variate may be regarded formally as a single analytical entity, denotable by a single letter (as Z), nevertheless it has two analytical constituents, or components: for instance, its real and imaginary constituents (X and Y); also, its modulus and amplitude ($|Z|$ and θ). Correspondingly, a complex variate can be represented geometrically by a single geometrical entity, namely a plane vector, but this, in turn, has two geometrical components, or constituents: for instance, its two rectangular components (X and Y); also, its two polar components, radius vector and vectorial angle ($R \equiv |Z|$ and θ).

This paper deals mainly with the modulus and the angle of the complex variate,[2] which are often of greater theoretical interest and practical im-

[1] "Probability Theory and Telephone Transmission Engineering," *Bell System Technical Journal*, January 1933, which will hereafter be referred to merely as the "1933 paper".

[2] Throughout the paper, I have used the term 'complex variate' for any 2-dimensional variate, because of the nature of the contemplated applications indicated in the first

portance than the real and imaginary constituents. The modulus variate and the angle variate, individually and jointly, are of considerable theoretical interest; while the modulus variate is also of very considerable practical importance, and the angle variate may conceivably become of some practical importance.

The paper is concerned chiefly with the 'distribution functions'[3] and the 'cumulative distribution functions' pertaining to the modulus (Sections 3 and 5) and to the angle (Sections 6 and 7) of the 'normal' complex variate, for the case where the mean value of this variate is zero. The distribution function for the reciprocal of the modulus is also included (Section 4).

The term 'probability function' is used in this paper generically to include 'distribution function' and 'cumulative distribution function.'

To avoid all except short digressions, some of the derivation work has been placed in appendices, of which there are four. These may be found of some intrinsic interest, besides facilitating the understanding of the paper.

1. Distribution Function and Cumulative Distribution Function in General: Definitions, Terminology, Notation, Relations, and Formulas

The present section constitutes a generic basis for the rest of the paper.

Let τ denote any complex variate, and let ρ and σ denote any pair of real quantities determining τ and determined by τ. (For instance, ρ and σ might be the real and imaginary components of τ, or they might be the modulus and angle of τ.) Geometrically, ρ and σ may be pictured as general curvilinear coordinates in a plane, as indicated by Fig. 1.1.

Let τ' denote the unknown value of a random sample consisting of a single τ-variate, and ρ' and σ' the corresponding unknown values of the constituents of τ'.

Further, let $G(\rho, \sigma)$ denote the 'areal probability density' at any point ρ,σ in the ρ,σ-plane, so that $G(\rho,\sigma)dA$ gives the probability that τ' falls in a differential area dA containing the point τ; and so that the integral of

---

paragraph of the Introduction, and also because the present paper is a sort of sequel to my 1933 paper, where the term 'complex variate' (or rather, 'complex chance-variable') was used throughout since there it seemed clearly to be the best term, on account of the field of applications contemplated and the specific applications given as illustrations. However, for wider usage the term 'bivariate' might be preferred because of its prevalence in the field of Mathematical Statistics; and therefore the paper should be read with this alternative in view.

[3] The term 'distribution function' is used with the same meaning in this paper as in my 1933 paper, although there the term 'probability law' was used much more frequently than 'distribution function,' but with the same meaning.

$G(\rho,\sigma)dA$ over the entire ρ,σ-plane is equal to unity, corresponding to certainty.

For the sake of subsequent needs of a formal nature, it will now be assumed that $G(\rho,\sigma) = 0$ at all points ρ,σ outside of the ρ_1, ρ_2, σ_1, σ_2 quadrilateral region in the ρ,σ-plane, Fig. 1.1, bounded by arcs of the four heavy curves, for which ρ has the values ρ_1 and ρ_2 and σ the values σ_1 and σ_2, with ρ_1 and σ_1 regarded, for convenience, as being less than ρ_2 and σ_2 respectively. Further, $G(\rho,\sigma)$ will be assumed to be continuous inside of this

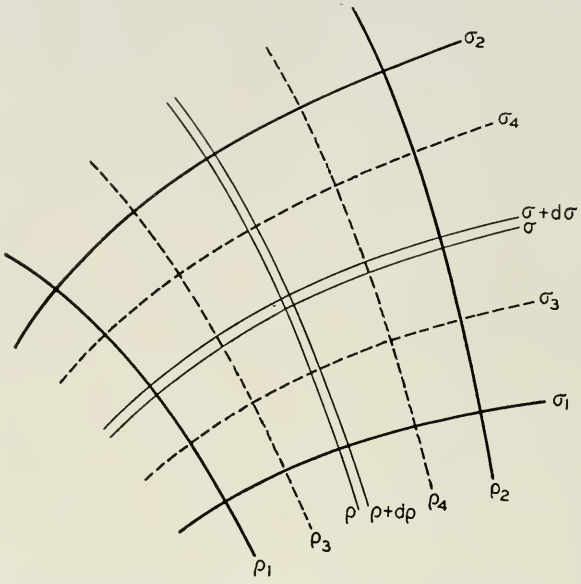


Fig. 1.1—Diagram of general curvilinear coordinates.

quadrilateral region, and to be non-infinite on its boundary. Hence, for probability purposes, it will suffice to deal with the open inequalities

$$\rho_1 < \rho < \rho_2, \qquad (1.1) \qquad\qquad \sigma_1 < \sigma < \sigma_2, \qquad (1.2)$$

which pertain to this quadrilateral region excluding its boundary; and thus it will not be necessary to deal with the closed inequalities $\rho_1 \leqq \rho \leqq \rho_2$ and $\sigma_1 \leqq \sigma \leqq \sigma_2$, which include the boundary.[4]

[4] The matters dealt with generically in this paragraph may be illustrated by the following two important particular cases, which occur further on, namely:

POLAR COORDINATES: $\rho = |\tau| = R$, $\sigma = \theta =$ angle of τ. Then $\rho_1 = R_1 = 0$, $\rho_2 = R_2 = \infty$, $\sigma_1 = \theta_1 = 0$, $\sigma_2 = \theta_2 = 2\pi$, whence (1.1) and (1.2) become $0 < R < \infty$ and $0 < \theta < 2\pi$, respectively.

RECTANGULAR COORDINATES: $\rho = \mathrm{Re}\,\tau = x$, $\sigma = \mathrm{Im}\,\tau = y$. Then $\rho_1 = x_1 = -\infty$, $\rho_2 = x_2 = \infty$, $\sigma_1 = y_1 = -\infty$, $\sigma_2 = y_2 = \infty$, whence (1.1) and (1.2) become $-\infty < x < \infty$ and $-\infty < y < \infty$, respectively.

A generic quadrilateral region contained within the quadrilateral region ρ_1, ρ_2, σ_1, σ_2 in Fig. 1.1 is the one bounded by arcs of the dashed curves ρ_3, ρ_4, σ_3, σ_4, where $\rho_3 < \rho_4$ and $\sigma_3 < \sigma_4$. Here, as in the preceding paragraph, it will evidently suffice to deal with open inequalities.

Referring to Fig. 1.1, the probability functions with which this paper will chiefly deal are certain particular cases of the probability functions $P(\rho, \sigma)$, $P(\rho \mid \sigma_{34})$ and $Q(\rho_{34}, \sigma_{34})$ occurring on the right sides of the following three equations respectively:

$$p(\rho < \rho' < \rho + d\rho, \sigma < \sigma' < \sigma + d\sigma) = P(\rho,\sigma)d\rho d\sigma, \tag{1.3}$$

$$p(\rho < \rho' < \rho + d\rho, \sigma_3 < \sigma' < \sigma_4) = P(\rho \mid \sigma_{34})d\rho, \tag{1.4}$$

$$p(\rho_3 < \rho' < \rho_4, \sigma_3 < \sigma' < \sigma_4) = Q(\rho_{34}, \sigma_{34}). \tag{1.5}$$

These equations serve to define the above-mentioned probability functions occurring on the right sides in terms of the probabilities denoted by the left sides, each expression $p(\ )$ on the left side denoting the probability of the pair of inequalities within the parentheses.[5] Inspection of these equations shows that: $P(\rho,\sigma)$ is the 'distribution function' for ρ and σ jointly; $P(\rho \mid \sigma_{34})$ is a 'distribution function' for ρ individually, with the understanding that σ' is restricted to the range σ_3-to-σ_4; $Q(\rho_{34},\sigma_{34})$ is a 'cumulative distribution function' for ρ and σ jointly.

Since the left sides of (1.3), (1.4) and (1.5) are necessarily positive, the right sides must be also. Hence, as all of the probability functions occurring in the right sides are of course desired to be positive, the differentials $d\rho$ and $d\sigma$ must be taken as positive, if we are to avoid writing $| d\rho |$ and $| d\sigma |$ in place of $d\rho$ and $d\sigma$ respectively.

Returning to (1.3), it is seen that, stated in words, $P(\rho,\sigma)$ is such that $P(\rho,\sigma)d\rho d\sigma$ gives the probability that the unknown values ρ' and σ' of the constituents of the unknown value τ' of a random sample consisting of a single τ-variate lie respectively in the differential intervals $d\rho$ and $d\sigma$ containing the constituent values ρ and σ respectively. Thus, unless $d\rho d\sigma$ is the differential element of area, $P(\rho,\sigma)$ is not equal to the 'areal probability density,' $G(\rho,\sigma)$, defined in the fourth paragraph of this section. In general, if E is such that $Ed\rho d\sigma$ is the differential element of area, then $P(\rho, \sigma) = EG(\rho, \sigma)$. (An illustration is afforded incidentally by Appendix A.)

$P(\rho,\sigma)$, defined by (1.3), is the basic 'probability function,' in the sense that the others can be expressed in terms of it, by integration. Thus

[5] Thus p in $p(\ )$ may be read 'probability that' or 'probability of.'

$P(\rho \mid \sigma_{34})$ and $P(\sigma \mid \rho_{34})$, defined respectively by (1.4) and by the correlative of (1.4), can be expressed as 'single integrals,' as follows[6]:

$$P(\rho \mid \sigma_{34}) = \int_{\sigma_3}^{\sigma_4} P(\rho,\sigma)\, d\sigma, \quad (1.6) \qquad P(\sigma \mid \rho_{34}) = \int_{\rho_3}^{\rho_4} P(\rho,\sigma)\, d\rho. \quad (1.7)$$

$Q(\rho_{34}, \sigma_{34})$, defined by (1.5), can be expressed as a 'double integral,' fundamentally; but, for purposes of analysis and of evaluation, this will be replaced by its two equivalent 'repeated integrals':

$$Q(\rho_{34}, \sigma_{34}) = \int_{\rho_3}^{\rho_4} \left[\int_{\sigma_3}^{\sigma_4} P(\rho,\sigma)\, d\sigma \right] d\rho = \int_{\sigma_3}^{\sigma_4} \left[\int_{\rho_3}^{\rho_4} P(\rho,\sigma)\, d\rho \right] d\sigma, \quad (1.8)$$

the set of integration limits being the same in both repeated integrals because these limits are constants, as indicated by Fig. 1.1. On account of (1.6) and (1.7) respectively, (1.8) can evidently be written formally as two single integrals:

$$Q(\rho_{34}, \sigma_{34}) = \int_{\rho_3}^{\rho_4} P(\rho \mid \sigma_{34})\, d\rho = \int_{\sigma_3}^{\sigma_4} P(\sigma \mid \rho_{34})\, d\sigma, \quad (1.9)$$

but implicitly these are repeated integrals unless the single integrations in (1.6) and (1.7) can be executed, in which case the integrals in (1.9) will actually be single integrals, and these will be quite unlike each other in form, being integrals with respect to ρ and σ respectively—though of course yielding a common expression in case the indicated integrations can be executed.

The particular cases of (1.4) and (1.5) with which this paper will chiefly deal are the following three:

$$p(\rho < \rho' < \rho + d\rho,\ \sigma_1 < \sigma' < \sigma_2) = P(\rho \mid \sigma_{12})\, d\rho \equiv P(\rho)\, d\rho, \quad (1.10)$$

$$p(\rho_1 < \rho' < \rho,\ \sigma_1 < \sigma' < \sigma_2) = Q(< \rho, \sigma_{12}) \equiv Q(\rho), \quad (1.11)$$

$$p(\rho < \rho' < \rho_2,\ \sigma_1 < \sigma' < \sigma_2) = Q(> \rho, \sigma_{12}) \equiv Q^*(\rho). \quad (1.12)$$

[6] The single-integral formulation in (1.6) can be written down directly by mere inspection of the left side of (1.4). Alternatively, (1.6) can be obtained by representing the left side of (1.4) by a repeated integral, as follows:

$$P(\rho \mid \sigma_{34})d\rho = \int_{\rho}^{\rho+d\rho} \left[\int_{\sigma_3}^{\sigma_4} P(\rho,\ \sigma)d\sigma \right] d\rho = \left[\int_{\sigma_3}^{\sigma_4} P(\rho,\ \sigma)d\sigma \right] d\rho,$$

whence (1.6); the last equality in the above chain equation in this footnote evidently results from the fact that, in general, $\int_x^{x+dx} f(x)dx = f(x)dx$, since each side of this equation represents dA, the differential element of area under the graph of $f(x)$ from x to $x + dx$.

In each of these three equations the very abbreviated notation at the extreme right will be used wherever the function is being dealt with extensively, as in the various succeeding sections. Such notation will not seem unduly abbreviated nor arbitrary if the following considerations are noted: In (1.10), σ_{12} corresponds to the entire effective range of σ, so that $P(\rho \mid \sigma_{12})$ is the 'principal' distribution function for ρ. Similarly, in (1.11), $Q(< \rho,\sigma_{12})$ is the 'principal' cumultive distribution function for ρ. In (1.12), the star indicates that $Q^*(\rho)$ is the 'complementary' cumulative distribution function, since $Q(\rho) + Q^*(\rho) = Q(\rho_{12}, \sigma_{12}) = 1$, unity being taken as the measure of certainty, of course.

For occasional use in succeeding sections, the defining equations for the probability functions pertaining to four other particular cases will be set down here:

$$p(\rho < \rho' < \rho + d\rho, \sigma_1 < \sigma' < \sigma) = P(\rho \mid < \sigma)\, d\rho, \tag{1.13}$$

$$p(\rho < \rho' < \rho + d\rho, \sigma < \sigma' < \sigma_2) = P(\rho \mid > \sigma)\, d\rho, \tag{1.14}$$

$$p(\rho_1 < \rho' < \rho, \sigma_1 < \sigma' < \sigma) = Q(< \rho, < \sigma), \tag{1.15}$$

$$p(\rho < \rho' < \rho_2, \sigma_1 < \sigma' < \sigma) = Q(> \rho, < \sigma). \tag{1.16}$$

It may be noted that (1.13) and (1.14) are mutually supplementary, in the sense that their sum is (1.10). Similarly, (1.15) and (1.16) are mutually supplementary, in the sense that their sum is $Q(\rho_{12}, < \sigma) = Q(< \sigma,\rho_{12})$, which is the correlative of (1.11).

This section will be concluded with the following three simple transformation relations (1.17), (1.18) and (1.19), which will be needed further on. They pertain to the probability functions on the right sides of equations (1.3), (1.4) and (1.5) respectively. h and k denote any positive real constants, the restriction to positive values serving to simplify matters without being too restrictive for the needs of this paper.

$$P(h\rho, k\sigma) = \frac{1}{hk} P(\rho,\sigma), \tag{1.17}$$

$$P(h\rho \mid k\sigma_{34}) = \frac{1}{h} P(\rho \mid \sigma_{34}), \tag{1.18}$$

$$Q(h\rho_{34}, k\sigma_{34}) = Q(\rho_{34}, \sigma_{34}). \tag{1.19}$$

Each of the three formulas (1.17), (1.18), (1.19) can be rather easily derived in at least two ways that are very different from each other. One way depends on probability inequality relations of the sort

$$p(t<t'<t+dt) = p(gt<gt'<gt+d[gt]), \tag{1.20}$$

$$p(t_3<t'<t_4) = p(gt_3<gt'<gt_4), \tag{1.21}$$

where t stands generically for ρ and for σ, and g is any positive real constant, standing generically for h and for k; (1.20) and (1.21) are easily seen to be true by imagining every variate in the universe of the t-variates to be multiplied by g, thereby obtaining a universe of (gt)-variates. A second way of deriving each of the three formulas (1.17), (1.18), (1.19) depends on general integral relations of the sort

$$\int_a^b f(t)\,dt = \frac{1}{g}\int_{ga}^{gb} f(t)\,d(gt) = \frac{1}{g}\int_{ga}^{gb} f\left(\frac{\lambda}{g}\right)d\lambda. \tag{1.22}$$

A third way, which is distantly related to the second way, depends on the use of the Jacobian for changing the variables in any double integral; thus,

$$\frac{P(\rho,\sigma)}{P(\lambda,\mu)} = \left|\frac{d\lambda d\mu}{d\rho d\sigma}\right| = \left|\frac{\partial(\lambda,\mu)}{\partial(\rho,\sigma)}\right| = 1 \div \left|\frac{\partial(\rho,\sigma)}{\partial(\lambda,\mu)}\right|, \tag{1.23}$$

the first equality in (1.23) depending on the fact that the two sets of variables and of differentials have corresponding values and hence are so related that

$$p(\rho<\rho'<\rho+d\rho,\ \sigma<\sigma'<\sigma+d\sigma) = p(\lambda<\lambda'<\lambda+d\lambda,\ \mu<\mu'<\mu+d\mu), \tag{1.24}$$

whence

$$P(\rho,\sigma)\,|\,d\rho d\sigma\,| = P(\lambda,\mu)\,|\,d\lambda d\mu\,|.$$

2. The Normal Complex Variate and Its Chief Probability Functions

The 'normal' complex variate may be defined in various equivalent ways. Here, a given complex variate $z = x + iy$ will be defined as being 'normal' if it is possible to choose in the plane of the scatter diagram of z a pair of rectangular axes, u and v, such that the distribution function[7] $P(u,v)$ for the given complex variate with respect to these axes can be written in the form[8]

$$P(u,v) = \frac{1}{2\pi S_u S_v}\exp\left[-\frac{u^2}{2S_u^2} - \frac{v^2}{2S_v^2}\right] = P(u)P(v). \tag{2.1}$$

We shall call $w = u + iv$ the 'modified' complex variate, as it represents the value of the given complex variate $z = x + iy$ when the latter is referred to the u,v-axes; $P(u)$ and $P(v)$ are respectively the individual distribution functions for the u and v components of the modified complex variate; and

[7] Defined by equation (1.3) on setting $\rho = u$ and $\sigma = v$.

[8] This equation is (12) of my 1933 paper. It can be easily verified that the (double) integral of (2.1) taken over the entire u, v-plane is equal to unity.

S_u and S_v are distribution parameters called the 'standard deviations' of u and v respectively. If t stands for u and for v generically, then

$$P(t) = \frac{1}{\sqrt{2\pi}S_t} \exp\left[-\frac{t^2}{2S_t^2}\right], \qquad (2.2) \qquad S_t^2 = \int_{-\infty}^{\infty} t^2 P(t)\, dt. \qquad (2.3)$$

From the viewpoint of the scatter diagram, the distribution function $P(u,v)$ is, in general, equal to the 'areal probability density' at the point u,v in the plane of the scatter diagram, so that the probability of falling in a differential element of area dA containing the point u,v is equal to $P(u,v)dA$; similarly, $P(u)$ and $P(v)$ are equal to the component probability densities. In particular, the probability density is 'normal' when $P(u,v)$ is given by (2.1).

Geometrically, equation (2.1) evidently represents a surface, the normal 'probability surface,' situated above the u, v-plane; and $P(u,v)$ is the ordinate from any point u,v in the u,v-plane to the probability surface.

The u,v-axes described above will be recognized as being the 'principal central axes,' namely that pair of rectangular axs which have their origin at the 'center' of the scatter diagram of $z = x + iy$ and hence at the center of the scatter diagram of $w = u + iv$, so that $\overline{w} = 0$, and are so oriented in the scatter diagram that $\overline{uv} = 0$ (whereas $\bar{z} \neq 0$ and $\overline{xy} \neq 0$, in general).

In equation (2.1), which has been adopted above as the analytical basis for defining the 'normal' complex variate, the distribution parameters are S_u and S_v ; and they occur symmetrically there, which is evidently natural and is desirable for purposes of definition. Henceforth, however, it will be preferable to adopt as the distribution parameters the quantities S and b defined by the pair of equations[9]

$$S^2 = S_u^2 + S_v^2, \qquad (2.4) \qquad\qquad bS^2 = S_u^2 - S_v^2, \qquad (2.5)$$

whence

$$b = \frac{S_u^2 - S_v^2}{S_u^2 + S_v^2} = \frac{1 - (S_v/S_u)^2}{1 + (S_v/S_u)^2}. \qquad (2.6)$$

From (2.4), S is seen to be a sort of 'resultant standard deviation.' The last form of (2.6) shows clearly that the total possible range of b is

$$-1 \leqq b \leqq 1, \qquad \text{corresponding to} \qquad \infty \geqq S_v/S_u \geqq 0.$$

The pair of simultaneous equations (2.4) and (2.5) give

$$2S_u^2 = (1+b)S^2, \qquad (2.7) \qquad\qquad 2S_v^2 = (1-b)S^2, \qquad (2.8)$$

which will be used below in deriving (2.11).

[9] Equations (2.4) and (2.6) are respectively (14) and (13) of my 1933 paper.

With the purpose of reducing the number of parameters by 1 and of dealing with variables that are dimensionless, we shall henceforth deal with the 'reduced' modified variate $W = U + iV$ defined by the equation

$$W = w/S = u/S + iv/S = U + iV. \tag{2.9}$$

Thus we shall be directly concerned with the scatter diagram of $W = U + iV$ instead of with that of $w = u + iv$.

The distribution function $P(U,V)$ for the rectangular components U and V of any complex variate $W = U + iV$ is defined by (1.3) on setting $\rho = U$ and $\sigma = V$; thus,

$$P(U,V)dUdV = p(U<U'<U+dU,\ V<V'<V+dV). \tag{2.10}$$

When the given variate $z = x + iy$ is normal, so that the modified variate $w = u + iv$ is normal, as represented by (2.1), then, since S is a mere constant, the reduced modified variate $W = U + iV$ defined by (2.9) will evidently be normal also, though of course with a different distribution parameter. Its distribution function $P(U,V)$ is found to have the formula[10]

$$P(U,V) = \frac{1}{\pi\sqrt{1-b^2}} \exp\left[-\frac{U^2}{1+b} - \frac{V^2}{1-b}\right] = P(U)P(V), \tag{2.11}$$

where $P(U)$ and $P(V)$ are the component distribution functions:

$$P(U) = \frac{1}{\sqrt{\pi(1+b)}} \exp\left[-\frac{U^2}{1+b}\right], \tag{2.12}$$

$$P(V) = \frac{1}{\sqrt{\pi(1-b)}} \exp\left[-\frac{V^2}{1-b}\right]. \tag{2.13}$$

These three distribution functions each contain only one distribution parameter, namely b; moreover, the variables $U = u/S$ and $V = v/S$ are dimensionless.

The distribution function $P(R,\theta)$ for the polar components R and θ of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by (1.3) on setting $\rho = R$ and $\sigma = \theta$; thus

$$P(R,\theta)dRd\theta = p(R<R'<R+dR,\ \theta<\theta'<\theta+d\theta). \tag{2.14}$$

For the case where W is 'normal,' it is shown in Appendix A that

$$P(R,\theta) = \frac{R}{\pi\sqrt{1-b^2}} \exp\left[\frac{-R^2}{1-b^2}(1 - b\cos 2\theta)\right] \tag{2.15}$$

$$= \frac{\sqrt{L}}{\pi} \exp[-L(1 - b\cos 2\theta)], \tag{2.16}$$

[10] This formula can be obtained from (2.1) by means of (2.7), (2.8), (2.9) and (1.17) after specializing (1.17) by the substitutions $\rho = u$, $\sigma = v$ and $h = k = 1/S$. It is (16) of my 1933 paper, but was given there without proof.

where

$$L = R^2/(1-b^2). \tag{2.17}$$

In $P(R,\theta)$ it will evidently suffice to deal with values of θ in the first quadrant, because of symmetry of the scatter diagram.

The fact that $P(R,\theta)$ depends on b as a parameter when W is 'normal' may be indicated explicitly by employing the fuller symbol $P(R,\theta;b)$ when desired; thus the former symbol is here an abbreviation for the latter.

In $P(R,\theta) \equiv P(R, \theta; b)$ it will suffice to deal with only positive values of b, that is, with $0 \leqq b \leqq 1$ (whereas the total possible range of b is $-1 \leqq b \leqq 1$). For (2.15) shows that changing b to $-b$ has the same effect as changing 2θ to $\pi \pm 2\theta$, or θ to $\pi/2 \pm \theta$; that is, $P(R,\theta; -b) = P(R, \pi/2 \pm \theta; b)$.

Seven formulas which will find considerable use subsequently are obtainable from the integrals corresponding to equations (1.13) to (1.16), by setting $\rho = R$ and $\sigma = \theta$ or else $\rho = \theta$ and $\sigma = R$, whichever is appropriate, and thereafter substituting for $P(R,\theta)$ the expression given by (2.16), and lastly executing the indicated integrations wherever they appear possible.[11] The resulting formulas are as follows:

$$P(R \mid < \theta) = \frac{\sqrt{L}}{\pi} \exp(-L) \int_0^\theta \exp(bL \cos 2\theta)\, d\theta, \tag{2.18}$$

$$P(\theta \mid < R) = \frac{\sqrt{1-b^2}}{2\pi} \frac{1 - \exp[-L(1 - b\cos 2\theta)]}{1 - b\cos 2\theta}, \tag{2.19}$$

$$P(\theta \mid > R) = \frac{\sqrt{1-b^2}}{2\pi} \frac{\exp[-L(1 - b\cos 2\theta)]}{1 - b\cos 2\theta}. \tag{2.20}$$

$$Q(< R, < \theta) = \frac{1}{\pi} \int_0^R \left[\sqrt{L} \exp(-L) \int_0^\theta \exp(bL\cos 2\theta)\, d\theta \right] dR \tag{2.21}$$

$$= \frac{\sqrt{1-b^2}}{2\pi} \int_0^\theta \frac{1 - \exp[-L(1 - b\cos 2\theta)]}{1 - b\cos 2\theta}\, d\theta, \tag{2.22}$$

$$Q(> R, < \theta) = \frac{1}{\pi} \int_R^\infty \left[\sqrt{L} \exp(-L) \int_0^\theta \exp(bL\cos 2\theta)\, d\theta \right] dR \tag{2.23}$$

$$= \frac{\sqrt{1-b^2}}{2\pi} \exp(-L) \int_0^\theta \frac{\exp(bL\cos 2\theta)}{1 - b\cos 2\theta}\, d\theta. \tag{2.24}$$

Formulas (2.21) to (2.24) are obtainable also by substituting (2.18) to (2.20) into the appropriate particular forms of (1.9).

When a θ-range of integration is 0-to-$q(\pi/2)$, where $q = 1, 2, 3$ or 4, this

[11] Except that in (2.22) the part $1/(1 - b\cos 2\theta)$ is integrable, as found in Sec. 7, equations (7.6) and (7.7).

range can be reduced to 0-to-$\pi/2$ provided the resulting integral is multiplied by q; that is,

$$\int_0^{q(\pi/2)} F(\theta)d\theta = q\int_0^{\pi/2} F(\theta)d\theta, \tag{2.25}$$

because of symmetry of the scatter diagram.

3. The Distribution Function for the Modulus

The distribution function $P(R \mid \theta_{12}) \equiv P(R)$ for the modulus R of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.10) on setting $\rho = R$, $\sigma = \theta$, $\sigma_1 = \theta_1 = 0$ and $\sigma_2 = \theta_2 = 2\pi$; thus

$$P(R)dR = p(R<R'<R+dR, \quad 0<\theta'<2\pi). \tag{3.1}$$

An integral formula for $P(R)$ is immediately obtainable from (1.6) by setting $\rho = R$, $\sigma = \theta$, $\sigma_3 = \sigma_1 = \theta_1 = 0$ and $\sigma_4 = \sigma_2 = \theta_2 = 2\pi$; thus

$$P(R) = \int_0^{2\pi} P(R,\theta)\, d\theta. \tag{3.2}$$

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $P(R)$ is here an abbreviation for $P(R; b)$. A formula for $P(R; b)$ can be obtained by substituting $P(R, \theta)$ from (2.15) into (3.2) and executing the indicated integration by means of the known Bessel function formula

$$\int_0^{\pi} \exp(\eta\cos\psi)\, d\psi = \pi I_0(\eta), \tag{3.3}$$

$I_0(\ )$ being the so-called 'modified Bessel function of the first kind,' of order zero.[12] The resulting formula is found to be[13]

$$P(R; b) = \frac{2R}{\sqrt{1-b^2}} \exp\left[\frac{-R^2}{1-b^2}\right] I_0\left[\frac{bR^2}{1-b^2}\right]. \tag{3.4}$$

This can also be obtained as a particular case of the more general formula (2.18) by setting $\theta = 2\pi$ in the upper limit of integration and then applying (3.3).

In $P(R;b)$ it will suffice to deal with positive values of b, that is, with $0 \leqq b \leqq 1$, as (3.4) shows that $P(R; -b) = P(R;b)$.

[12] It may be recalled that $I_0(z) = J_0(iz)$, and in general that $I_n(z) = i^{-n}J_n(iz)$.
In the list of references on Bessel functions, on the last page of this paper, the 'modified Bessel function' is treated in Ref. 2, p. 20; Ref. 3, p. 102; Ref. 4, p. 41; Ref. 1, p. 77.
Regarding formula (3.3), see Ref. 1, p. 181, Eq. (4), $\nu = 0$; Ref. 1, p. 19, Eq. (9), fourth expression, $\nu = 0$; Ref. 2, p. 46, Eq. (10), $n = 0$; Ref. 3, p. 164, Eq. 103, $n = 0$.

[13] This formula was given in its cumulative forms, $\int P(R; b)dR$, as formulas (51-A) and (53-A) of the unpublished Appendix A to my 1933 paper.

It will often be advantageous to express $P_{R;b}$ in terms of b and one or the other of the auxiliary variables L and T defined by the equations

$$L = \frac{R^2}{1 - b^2}, \quad (3.5) \qquad T = bL = \frac{bR^2}{1 - b^2}. \quad (3.6)$$

Formula (3.4) thereby becomes, respectively,

$$P(R;b) = 2\sqrt{L}\exp(-L)I_0(bL), \quad (3.7)$$

$$P(R;b) = 2\sqrt{\frac{T}{b}}\exp\left[\frac{-T}{b}\right]I_0(T). \quad (3.8)$$

Formula (3.8) will often be preferable to (3.7) because the argument of the Bessel function in (3.8) is a single quantity, T.

Because tables of $I_0(X)$ are much less easily interpolated than tables of $M_0(X)$ defined by the equation

$$M_0(X) = \exp(-X)I_0(X), \quad (3.9)$$

extensive tables of which have been published,[14] it is natural, at least for computational purposes, to write (3.4) in the form

$$P(R;b) = \frac{2R}{\sqrt{1 - b^2}}\exp\left[\frac{-R^2}{1 + b}\right]M_0\left[\frac{bR^2}{1 - b^2}\right]. \quad (3.10)$$

For use in equation (3.16), it is convenient to define here a function $M_1(X)$ by the equation

$$M_1(X) = \exp(-X)I_1(X), \quad (3.11)$$

corresponding to (3.9) defining $M_0(X)$. $M_1(X)$ has the similar property that it is much more easily interpolated than is $I_1(X)$; and extensive tables of $M_1(X)$ are constituent parts of the tables in Ref. 1 and Ref. 6.

The quantity $bR^2/(1-b^2) \equiv T$, which occurs in (3.4) and (3.8) as the argument of $I_0(\ )$, and in (3.10) as the argument of $M_0(\ )$, evidently ranges from 0 to ∞ when R ranges from 0 to ∞ and also when b ranges from 0 to 1. Formula (3.10) is suitable for computational purposes for all values of the above-mentioned argument $bR^2/(1-b^2) \equiv T$ not exceeding the largest values of X in the above-cited tables in Ref. 1 and Ref. 6. For larger values of the argument, and partiularly for dealing with the limiting

[14] Ref. 1, Table II (p. 698–713), for $X = 0$ to 16 by .02. Ref. 6, Table VIII (p. 272–283), for $X = 5$ to 10 by .01, and 10 to 20 by 0.1. Each of these references conveniently includes a table of $\exp(X)$ whereby values of $I_0(X)$ can be readily and accurately evaluated if desired. Values of $I_0(X)$ so obtained would enable formulas (3.4), (3.7) and (3.8) of the present paper to be used with high accuracy without any difficult interpolations, since the table of $\exp(X)$ is easily interpolated by utilizing the identity $\exp(X_1 + X_2) = \exp(X_1)\exp(X_2)$.

case where the argument becomes infinite, formula (310)—and hence (3.4)—may be advantageously written in the form

$$P(R;\, b) = \frac{2}{\sqrt{2\pi b}} \exp\left[\frac{-R^2}{1+b}\right] N_0\left[\frac{bR^2}{1-b^2}\right], \tag{3.12}$$

where

$$N_0(X) = \sqrt{2\pi X}\, \exp(-X) I_0(X) = \sqrt{2\pi X}\, M_0(X), \tag{3.13}$$

an extensive table of which has been published.[15] The natural suitability of the function $N_0(X)$ for dealing with large values of X is evident from the structure of the asymptotic series for $N_0(X)$, for sufficiently large values of X, which runs as follows:[16]

$$N_0(X) \sim 1 + \frac{1^2}{1!8X} + \frac{1^2 3^2}{2!(8X)^2} + \frac{1^2 3^2 5^2}{3!(8X)^3} + \cdots, \tag{3.14}$$

whence it is evident that

$$N_0(\infty) = 1. \tag{3.15}$$

For use in Appendix C, it is convenient to define here a function $N_1(X)$ by the equation[17]

$$N_1(X) = \sqrt{2\pi X}\, \exp(-X) I_1(X) = \sqrt{2\pi X} M_1(X), \tag{3.16}$$

corresponding to (3.13) defining $N_0(X)$, with $M_1(X)$ defined by (3.11). The asymptotic series for $N_1(X)$, which will be needed in Appendix C, is[18]

$$N_1(X) \sim 1 - 3\left[\frac{1}{1!8X} + \frac{(1\cdot 5)}{2!(8X)^2} + \frac{(1\cdot 5)(3\cdot 7)}{3!(8X)^3} + \cdots\right], \tag{3.17}$$

whence it is evident that

$$N_1(\infty) = 1. \tag{3.18}$$

When b is very nearly but not exactly equal to unity, so that

$$\frac{bR^2}{1-b^2} \approx \frac{R^2}{1-b^2} \approx \frac{R^2}{2(1-b)}, \tag{3.19}$$

it is seen from (3.4) that $P(R;b)$ is, to a very close approximation, a function

[15] Ref. 7, pp. 45–72, for $X = 10$ to 50 by 0.1, 50 to 200 by 1, 200 to 1000 by 10, and for various larger values of X.

[16] Ref. 1, p. 203, with (ν, m) defined on p. 198; Ref. 5, p. 366; Ref. 2, p. 58; Ref. 3, p. 163, Eq. 84; Ref. 4, pp. 48, 84.

[17] $N_1(X)$ is tabulated along with $N_0(X)$ in Ref. 7 already cited in connection with equation (3.13).

[18] Ref. 1, p. 203, with (ν, m) defined on p. 198; Ref. 5, p. 366; Ref. 2, p. 58; Ref. 3, p. 163, Eq. 84.

of only a single quantity, which may be any one of the three very nearly equal expressions in (3.19)—but the last of them is evidently the simplest.

Fig. 3.1 gives curves of $P(R;b)$, with the variable R ranging continuously

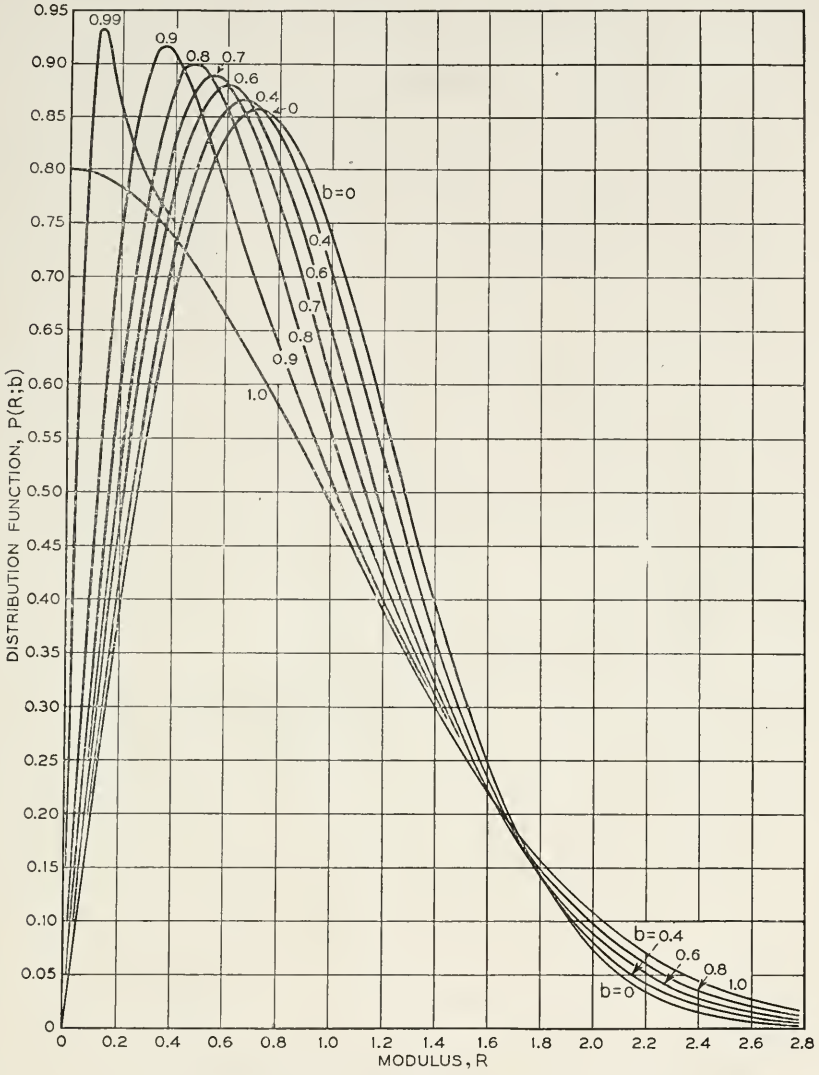


Fig. 3.1—Distribution function for the modulus ($R = 0$ to 2.8).

from 0 to 2.8 and the parameter b ranging by steps from 0 to 1 inclusive, which is the complete range of positive b. Fig. 3.2 gives an enlargement (along the R-axis) of the portion of Fig. 3.1 between $R = 0$ and $R = 0.4$,

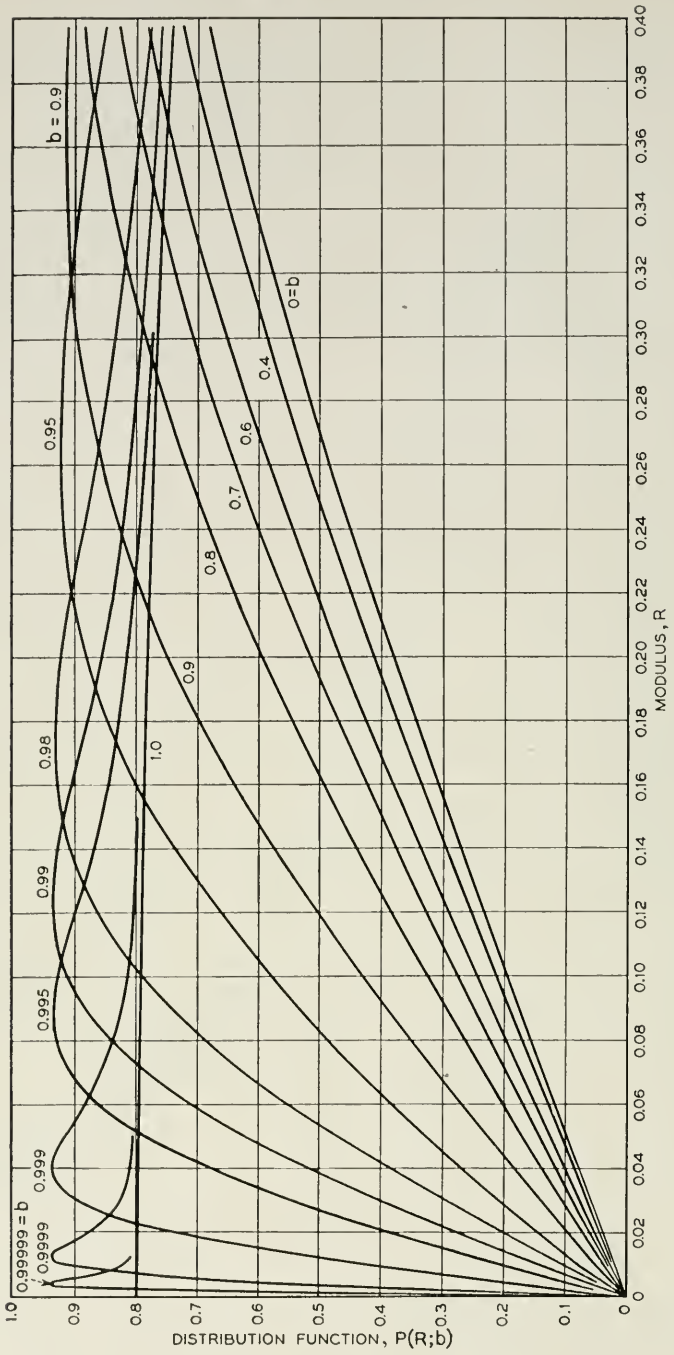


Fig. 3.2—Distribution function for the modulus (R = 0 to 0.4).

and includes therein curves for a considerable number of additional values of b between 0.9 and 1 so chosen as to show clearly how, with b increasing toward 1, the curves approach the curve for $b = 1$ as a limiting particular curve; or, conversely, how the curve for $b = 1$ constitutes a limiting particular curve—which, incidentally, will be found to be a natural and convenient reference curve. This curve, for $b = 1$, will be considered more fully a little further on, because it is a limiting particular curve and because of its resulting peculiarity at $R = 0$, the curve for $b = 1$ having at $R = 0$ a projection, or spur, situated in the $P(R;b)$ axis and extending from 0.7979 to 0.9376 therein (as shown a little further on).

The formulas and curves for $b = 0$ and $b = 1$, being of especial interest and importance, will be considered before the remaining curves of the set.

For the case $b = 0$, formula (3.4) evidently reduces immediately to

$$P(R;0) = 2R \exp(-R^2). \tag{3.20}$$

This case, $b = 0$, is that degenerate particular case in which the equiprobability curves in the scatter diagram of the complex variate, instead of being ellipses (concentric), are merely circles, as noted in my 1933 paper, near the bottom of p. 44 thereof (p. 10 of reprint).

For the case $b = 1$, the formula for the entire curve of $P(R; b) = P(R;1)$, except only the part at $R = 0$, can be obtained by merely setting $b = 1$ in[19] (3.12) as this, on account of (3.15), thereby reduces immediately to

$$P'(R; 1) = \frac{2}{\sqrt{2\pi}} \exp\left[-\frac{R^2}{2}\right], \qquad (R \neq 0), \tag{3.21}$$

$P'(R;1)$ denoting the value of $P(R;b)$ when $b = 1$ but $R \neq 0$, the restriction $R \neq 0$ being necessary because the quantity $R^2/(1-b^2)$ in (3.12)—and in (3.4)—does not have a definite value when $b = 1$ if $R = 0$. Thus, in Figs. 3.1 and 3.2, the curve of $P'(R;1)$ is that part of the curve for $b = 1$ which does not include any point in the $P(R; b)$ axis (where $R = 0$) but extends rightward from that axis toward $R = +\infty$. The curve of $P'(R;1)$ is the 'effective' part of the curve of $P(R;1)$, in the sense that the area under the former is equal to that under the latter, since the part of the curve of $P(R;1)$ at $R = 0$ can have no area under it.

$P(0;1)$ denoting (by convention) the value, or values, of $P(R;b)$ when $R = 0$ and $b = 1$, that is, the value, or values, of $P(R;1)$ when $R = 0$, it is seen, from consideration of the curves of $P(R;b)$ in Figs. 3.1 and 3.2 when b approaches 1 and ultimately becomes equal to 1, that the curve of $P(0;1)$ consists of all points in the vertical straight line segment extending upward in the $P(R;b)$ axis, from the origin to a height 0.9376 [= Max $P(R;1)$],[20]

[19] Use of (3.12) instead of (3.4), which is transformable into (3.12), avoids the indefinite expression $\infty .0. \infty$ which would result directly from setting $b = 1$ in (3.4).

[20] As shown near the end of Appendix B, Max $P(R;1)$ is situated at $R = 0$ and is equal to 0.9376.

together with all points in the straight line segment extending downward from the point at 0.9376 to the point at 0.7979 [$= 2/\sqrt{2\pi} = P'(R;1)$ for $R = 0+$]. The curve of $P(0;1)$, because it has no area under it, is the 'non-effective' part of the curve of $P(R;1)$.

Starting at the origin of coordinates, where $R = 0$, the complete curve of $P(R;1)$ consists of the curve of $P(0;1)$, described in the preceding paragraph, in sequence with the curve of $P'(R;1)$, given by (3.21). Thus the complete curve of $P(R;1)$ is the locus of a tracing point moving as follows: Starting at the origin of coordinates, the tracing point first ascends in the $P(R;\ b)$ axis to a height 0.9376 [$=$ Max $P(R;1)$]; second, descends from 0.9376 to 0.7979 [$= 2/\sqrt{2\pi} = P'(R;1)$ for $R = 0+$]; and, third, moves rightward along the graph of $P'(R;1)$ [$b = 1$] toward $R = +\infty$. The locus of all of the points thus traversed by the tracing point is the complete curve[21] of $P(R;1)$.

In addition to being the principal part ('effective' part) of the curve of $P(R;1)$, the curve of $P'(R;1)$, whose formula is (3.21), has a further important significance. For the right side of (3.21), except for the factor 2, will be recognized as being the expression for the well-known 1-dimensional 'normal' law; the presence of the factor 2 is accounted for by the fact that the variable $R = |\ R\ |$ can have only posiive values and yet the area under the curve must be equal to unity. This case, $b = 1$, is that degenerate particular case in which the equiprobability curves, instead of being ellipses, are superposed straight line segments, so that the resulting 'probability density' is not constant but varies in accordance with the 1-dimensional 'normal' law (for real variates), as noted in my 1933 paper, at the top of p. 45 thereof (p. 11 of reprint).

All of the curves of $P(R;b)$, where $0 \leqq b \leqq 1$, pass through the origin, the curve of $P(R;1)$ [$b = 1$] being no exception, since the part $P(0;1)$ passes through the origin.

Formula (3.12), supplemented by (3.15), shows that $P(R;\ b) = 0$ at $R = \infty$; and this is in accord with the consideration that the total area under the curve of $P(R;b)$ must be finite (equal to unity).

Since $P(R;b) = 0$ at $R = 0$ and at $R = \infty$, every curve of $P(R;b)$ must have a maximum value situated somewhere between $R = 0$ and $R = \infty$— as confirmed by Figs. 3.1 and 3.2. These figures show that when b increases from 0 to 1 the maximum value increases throughout but the value of R where it is located decreases throughout.

The maxima of the function $P(R;b)$ and of its curves (Figs. 3.1 and 3.2) are of considerable theoretical interest and of some practical importance.

[21] The presence, in the curve of $P(R;\ 1)$, of the vertical projection, or spur, situated in the $P(R;\ b)$ axis and extending from 0.7979 to 0.9376 therein, is somewhat remindful (qualitatively) of the 'Gibbs phenomenon' in the representation of discontinuous periodic functions by Fourier series.

The cases $b = 0$ and $b = 1$ will be dealt with first, and then the general case ($b = b$).

For the case $b = 0$ it is easily found by differentiating (3.20) that $P(R;b) = P(R;0)$ is a maximum at $R = 1/\sqrt{2} = 0.7071$ and hence that its maximum value is $\sqrt{2}\exp(-1/2) = 0.8578$, agreeing with the curve for $b = 0$ in Fig. 3.1.

For the case $b = 1$, which is a limiting particular case, the maximum value of $P(R;b) = P(R;1)$ apparently cannot be found driectly and simply, as will be realized from the preceding discussion of this case. Near the end of Appendix B, it is shown that the maximum value of $P(R;1)$ occurs at $R = 0$ (as would be expected) and is equal to 0.9376. This is the maximum value of the part $P(0;1$ of $P(R;1)$. The remaining part of $P(R;1)$, namely $P'(R;1)$, whose formula is (3.21), is seen from direct inspection of that formula to have a right-hand maximum value at $R = 0+$, whence this maximum value is $2/\sqrt{2\pi} = 0.7979$.

For the general case when b has any fixed value within its possible positive range ($0 \leqq b \leqq 1$), it is apparently not possible to obtain an explicit expression (in closed form) either for the value of R at which $P(R;b)$ has its maximum value or for the maximum value of $P(R;b)$; and hence it is not possible to make explicit computations of these quantities for use in plotting curves of them, versus b, of which they will evidently be functions. However, as shown in Appendix B, these desired curves can be exactly computed, in an indirect manner, by temporarily taking b as the dependent variable and taking T, defined by (3.6), as an intermediate independent variable. For let R_C denote the critical value of R, that is, the value of R at which $P(R;b)$ has its maximum value; and let T_C denote the corresponding value of T, whence, by (3.6),

$$T_C = bR_C^2/(1-b^2). \tag{3.22}$$

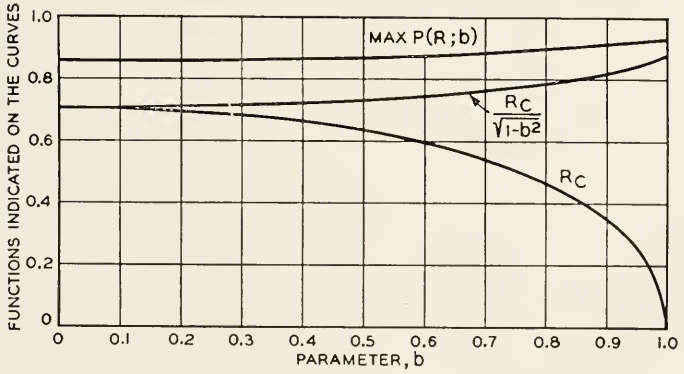


Fig. 3.3—Functions relating to the maxima of the distribution function for the modulus.

Then, computed by means of the formulas derived in Appendix B, Fig. 3.3 gives a curve of R_C and a curve of Max $P(R;b)$, each versus b. Since the curve of R_C cannot be read accurately at $b \approx 1$, there is included also a curve of $R_C/\sqrt{1 - b^2}$, from which R_C can be accurately and easily computed for any value of b; incidentally, the curve of $R_C/\sqrt{1 - b^2}$ is simultaneously a curve of $\sqrt{T_C/b}$, on account of (3.22). From Fig. 3.3 it is seen that R_C varies greatly with b but that Max $P_{R;b}$ varies only a little, as also is seen from inspection of Figs. 3.1 and 3.2 giving curves of $P(R;b)$ as function of R with b as parameter.

In Fig. 3.3, the curve of R_C shows that for $b = 1$ the maximum of $P(R;b)$ occurs at $R = 0$; and the curve of Max $P(R;b)$ shows that Max $P(R;1) \approx 0.94$, agreeing to two significant figures with the value 0.9376 found near the end of Appendix B.

4. The Distribution Function for the Reciprocal of the Modulus

At first, let R denote any real variate, and $P(R)$ its distribution function. Also let r denote the reciprocal of R, so that $r = 1/R$; and let $P(r)$ denote the distribution function for r. Then [22]

$$P(r) = R^2 P(R) = P(R)/r^2. \tag{4.1}$$

If $P(R)$ depends on any parameters, $P(r)$ will evidently depend on the same parameters.

The rest of this section deals with the case where $W \equiv R(\cos\theta + i \sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $P(R)$ and $P(r)$ are here abbreviations for $P(R;b)$ and $P(r;b)$ respectively.

As $P(R;b)$ has the distribution function given by (3.4), the distribution function for r will be

$$P(r;b) = \frac{2}{(\sqrt{1 - b^2})r^3} \exp\left[\frac{-1}{(1 - b^2)r^2}\right] I_0\left[\frac{b}{(1 - b^2)r^2}\right], \tag{4.2}$$

obtained from the right side of (3.4) by changing R to $1/r$ and multiplying

[22] For if r and R denote any two real variates that are functionally related, say $F(r, R) = 0$, and if dr and dR are corresponding small increments, then evidently

$$P(r)\,|\,dr\,| = P(R)\,|\,dR\,| \quad \text{whence} \quad \frac{P(r)}{P(R)} = \left|\frac{dR}{dr}\right| = \left|\frac{\partial F/\partial r}{\partial F/\partial R}\right|.$$

In particular, if $r = 1/R$, whence $F = r - 1/R$, then (4.1) results immediately.

For a somewhat different and more detailed treatment of change of the variable in distribution functions, see Thorton C. Fry, "Probability and its Engineering Uses," 1928, pp. 153–155. (Cases of more than one variate are treated on pp. 155–174 of the same reference.)

the result by $1/r^2$, in accordance with (4.1). Evidently $P(r;-b) = P(r;b)$.

By means of (4.1), formulas (3.7) and (3.8) give, respectively,

$$P(r;b) = 2(1-b^2)L^{3/2}\exp(-L)I_0(bL), \tag{4.3}$$

$$P(r;b) = 2(1-b^2)\left[\frac{T}{b}\right]^{3/2}\exp\left[\frac{-T}{b}\right]I_0(T), \tag{4.4}$$

wherein L and T are defined by (3.5) and (3.6) respectively, but will now be written in the equivalent forms

$$L = \frac{1}{(1-b^2)r^2}, \tag{4.5}$$

$$T = bL = \frac{b}{(1-b^2)r^2}, \tag{4.6}$$

which are evidently more suitable for the present section.

A few particular cases that are especially important will be dealt with in the following brief paragraph, ending with equation (4.8).

For the two extreme values of r, namely 0 and ∞, $P(r;b)$ is zero for all values of b in the b-range $(0 \leqq b \leqq 1)$.

When $b = 0$,

$$P(r;b) = P(r;0) = \frac{2}{r^3}\exp\left[\frac{-1}{r^2}\right]. \tag{4.7}$$

When $b = 1$,

$$P(r;b) = P(r;1) = \frac{2}{\sqrt{2\pi}}\frac{1}{r^2}\exp\left[\frac{-1}{2r^2}\right]. \tag{4.8}$$

Fig. 4.1 gives curves of $P(r;b)$, with the variable r ranging continuously from 0 to 1.4 and the parameter b ranging by steps from 0 to 1; however, in the r-range where r is less than about 0.6, alternate curves had to be omitted to avoid undue crowding. Fig. 4.2 gives an enlargement of the section betwen $r = 0.2$ and $r = 0.5$, and includes therein the curves that had to be omitted from Fig. 4.1.

In Fig. 4.1 it will be noted that with the scale there used for $P(r;b)$ the values of $P(r;b)$ are too small to be even detectable for values of r less than about 0.25. Even in the enlargement supplied by Fig. 4.2, the values of $P(r;b)$ are not detectable for r less than about 0.2.

The curves of $P(r;b)$ in Figs. 4.1 and 4.2 would have had to be computed from the lengthy formula (4.2)—or its equivalents—except for the fact that curves of $P(R;b)$ had already been computed in the preceding section of the paper. The last circumstance enabled the $P(r;b)$ curves to be obtained from the $P(R;b)$ curves by means of the very simple relation (4.1).

It will be observed that each curve of $P(r;b)$ [Fig. 4.1] has a maximum

ordinate, whose value and location depend on b. When b increases from 0 to 1, the maximum ordinate decreases throughout but the value of r where it is located remains nearly constant, at about 0.82, until b becomes about

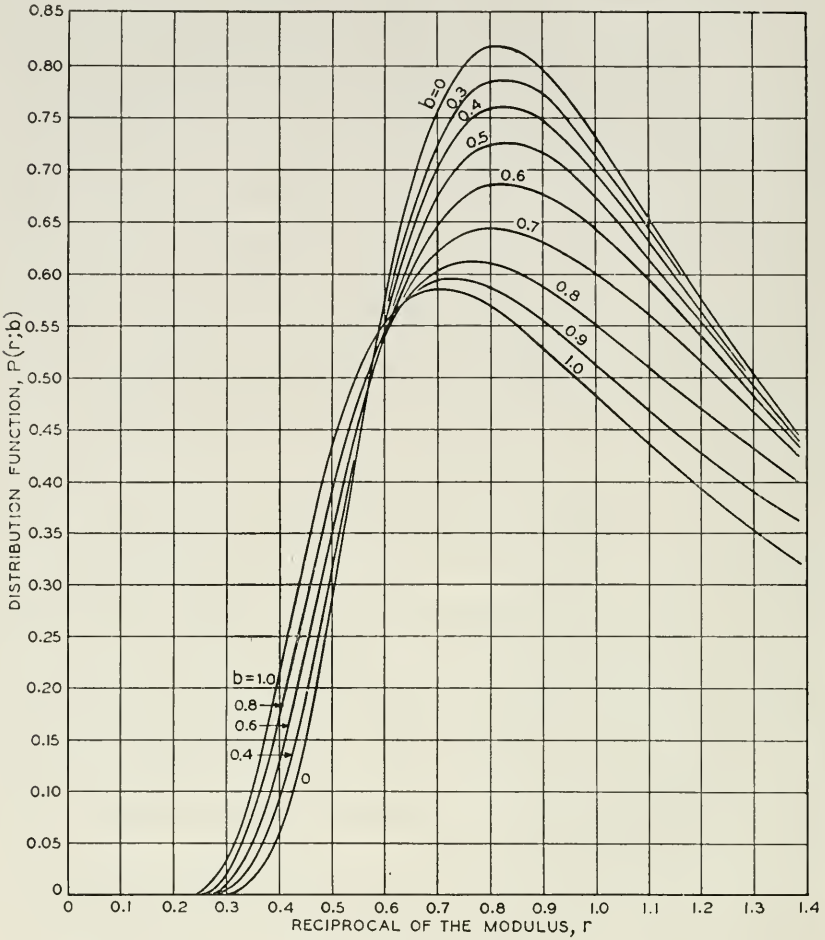


Fig. 4.1—Distribution function for the reciprocal of the modulus ($r = 0$ to 1.4).

0.7, after which the location of the maximum value moves rather rapidly to about 0.71 for $b = 1$.

For the cases $b = 0$ and $b = 1$, it is easily found, by differentiating (4.7) and (4.8), that the maximum ordinates are located at $r = \sqrt{2/3} = 0.8165$ and at $r = 1/\sqrt{2} = 0.7071$ respectively; and hence, by (4.7) and (4.8), that the values of these maximum ordinates are $(3\sqrt{3/2}\exp(-3/2) =$

0.8198 and $(4/\sqrt{2\pi})\exp(-1) = 0.5871$ respectively. These results for the cases $b = 0$ and $b = 1$ agree with the corresponding curves in Fig. 4.1.

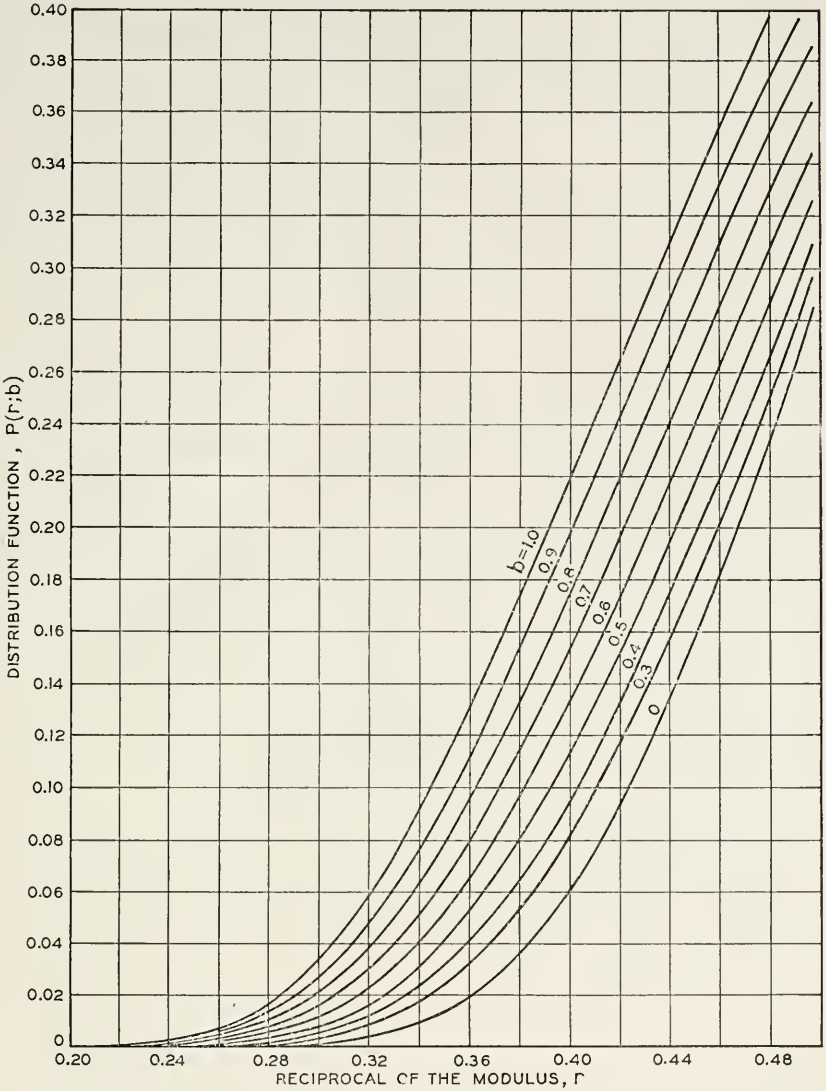


Fig. 4.2—Distribution function for the reciprocal of the modulus ($r = 0.2$ to 0.5).

For the general case where b has any fixed value in the b-range $(0 \leqq b \leqq 1)$, it is apparently not possible to obtain an explicit expression (in closed form) either for the value of r at which $P(r;b)$ has its maximum value or for the

maximum value of $P(r;b)$. However, as shown in Appendix C, curves of these quantities versus b can be computed, in an indirect manner, by temporarily taking b as the dependent variable and taking T, defined by (4.6), as an intermediate independent variable. For let r_c denote the critical value of r, that is, the value of r at which $P(r;b)$ has its maximum value; and let T_c denote the corresponding value of T, whence, by (4.6),

$$T_c = b/(1-b^2)r_c^2. \tag{4.9}$$

Then, computed by means of the formulas derived in Appendix C, Fig. 4.3 gives a curve of r_c and a curve of Max $P(r;b)$, each versus b. From these curves it is seen that r_c and Max $P(r;b)$ do not vary greatly with b, as also is seen from inspection of Fig. 4.1 giving curves of $P(r;b)$ as function of r with b as parameter.

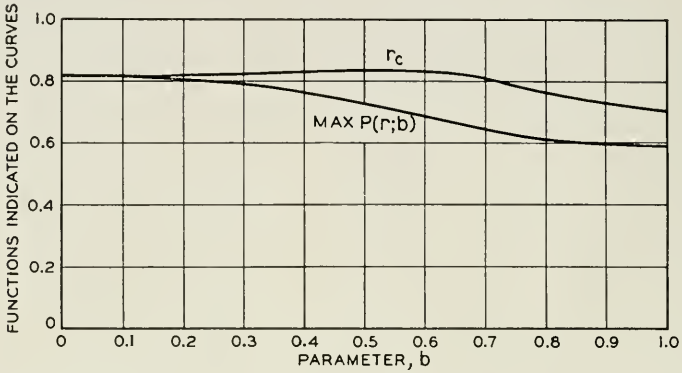


Fig. 4.3—Functions relating to the maxima of the distribution function for the reciprocal of the modulus.

5. The Cumulative Distribution Function for the Modulus

The cumulative distribution function $Q(<R,\theta_{12}) \equiv Q(R)$ for the modulus R of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.11) on setting $\rho = R$, $\sigma = \theta$, $\rho_1 = R_1 = 0$, $\sigma_1 = \theta_1 = 0$ and $\sigma_2 = \theta_2 = 2\pi$; thus

$$Q(R) = p(0<R'<R,\ 0<\theta'<2\pi). \tag{5.1}$$

Similarly, from (1.12), the complementary cumulative distribution function $Q(>R,\theta_{12}) \equiv Q^*(R)$ is defined by the equation

$$Q^*(R) = p(R<R'<\infty,\ 0<\theta'<2\pi). \tag{5.2}$$

$Q^*(R)$ is usually more convenient than $Q(R)$ for use in engineering applications, because it is usually more convenient to deal with the relatively

small probability of exceeding a preassigned rather large value of R than to deal with the corresponding rather large probability (nearly equal to unity) of being less than the preassigned value of R.

A 'double integral' for $Q(R)$, in the form of two 'repeated integrals,' can be written down directly by inspection of the $p(\ )$ expression in (5.1) or by specialization of (1.8); thus

$$Q(R) = \int_0^R \left[\int_0^{2\pi} P(R,\theta)\, d\theta \right] dR = \int_0^{2\pi} \left[\int_0^R P(R,\theta)\, dR \right] d\theta. \qquad (5.3)$$

Evidently these can be written formally as two 'single integrals,'

$$Q(R) = \int_0^R P(R)\, dR = \int_0^{2\pi} P(\theta \mid < R)\, d\theta, \qquad (5.4)$$

by means of the distribution functions $P(R) = P(R \mid \theta_{12})$ and $P(\theta \mid < R)$ given by the formulas

$$P(R) = \int_0^{2\pi} P(R,\theta)\, d\theta, \qquad (5.5) \qquad P(\theta \mid < R) = \int_0^R P(R,\theta)\, dR. \qquad (5.6)$$

(5.5) is the same as (3.2). (5.6) is a special case of (1.6), and the left side of (5.6) is a special case of $P(\rho \mid < \sigma)$ defined by (1.13).

Similarly, from (5.2), we arrive at the following formulas corresponding to (5.3), (5.4), (5.5), and (5.6) respectively:

$$Q^*(R) = \int_R^\infty \left[\int_0^{2\pi} P(R,\theta)\, d\theta \right] dR = \int_0^{2\pi} \left[\int_R^\infty P(R,\theta)\, dR \right] d\theta, \qquad (5.7)$$

$$Q^*(R) = \int_R^\infty P(R)\, dR = \int_0^{2\pi} P(\theta \mid > R)\, d\theta, \qquad (5.8)$$

$$P(R) = \int_0^{2\pi} P(R,\theta)\, d\theta, \qquad (5.9) \qquad P(\theta \mid > R) = \int_R^\infty P(R,\theta)\, dR. \qquad (5.10)$$

The rest of this section deals with the case where $W \equiv R(\cos\theta + i \sin\theta)$ is 'normal.'[23] Since this case depends on b as a parameter, $Q(R)$ and $Q^*(R)$ are here abbreviations for $Q(R;b)$ and $Q^*(R;b)$ respectively.

A natural and convenient way for deriving formulas for $Q(R)$ is afforded by the general formula (5.4) together with the auxiliary general formulas (5.5) and (5.6), beginning with the two latter.

For the 'normal' case, $P(R,\theta)$ is given by (2.15). When this is substituted into (5.5) and (5.6), it is found that each of the indicated integra-

[23] For the 'normal' case, the cumulative distribution function was treated in a very different manner in my 1933 paper and its unpublished Appendix A. That paper included applications to two important practical problems, and its unpublished Appendix C treated a third such problem. (The unpublished appendices, A, B and C, are mentioned in footnote 3 of the 1933 paper.)

tions can be executed, giving the two previously obtained formulas (3.4) and (2.19) for $P(R) \equiv P(R;b)$ and $P(\theta \mid <R)$ respectively. When these are substituted into (5.4), there result two types of single-integral formulas for $Q(R)$: A primary type, involving an indicated integration as to R; and a secondary type, involving an indicated integration as to θ. Formulas of these two types for $Q(R)$ will now be derived.

An integral formula of the primary type for $Q(R) \equiv Q(R;b)$ can be obtained by substituting $P(R) \equiv P(R;b)$ from (3.4) into the first integral in (5.4), giving

$$Q(R) = 2\int_0^R \frac{\lambda}{\sqrt{1-b^2}} \exp\left[\frac{-\lambda^2}{1-b^2}\right] I_0\left[\frac{b\lambda^2}{1-b^2}\right] d\lambda. \qquad (5.11)$$

This can also be obtained as a particular case of the more general formula (2.21) by setting $\theta = 2\pi$ in the upper limit of integration and then applying (3.3).

In (5.11), λ is used instead of R as the integration variable in order to avoid any possible confusion with R as an integration limit. Thus the integrand is a function of λ with b as a parameter. Evidently $Q(R;b) = Q(R;-b)$. Formula (5.11) is evidently suitable for evaluation of $Q(R)$ by numerical integration.[24]

By suitably changing the variable in (5.11), we arrive at the following various additional formulas, which, though equivalent to (5.11), are very different as regards the integrand and the limits of integration. As previously, L denotes $R^2/(1-b^2)$.

$$Q(R) = \frac{1}{\sqrt{1-b^2}} \int_0^{R^2} \exp\left[\frac{-\lambda}{1-b^2}\right] I_0\left[\frac{b\lambda}{1-b^2}\right] d\lambda, \qquad (5.12)$$

$$Q(R) = \sqrt{1-b^2} \int_0^L \exp(-\lambda)\ I_0(b\lambda)\ d\lambda, \qquad (5.13)$$

$$Q(R) = L\sqrt{1-b^2} \int_0^1 \exp(-L\lambda)\ I_0(bL\lambda)\ d\lambda, \qquad (5.14)$$

$$Q(R) = \sqrt{1-b^2} \int_{\exp(-L)}^1 I_0(b \log \lambda)\ d\lambda. \qquad (5.15)$$

These four additional formulas are of some theoretical interest, but apparently they are less suitable than (5.11) for numerical integration with respect to R. A formula differing slightly from (5.11) could evidently be obtained by taking $\lambda/\sqrt{1-b^2}$ as a new variable, and hence $R/\sqrt{1-b^2}$ as the upper limit of integration.

Corresponding formulas for $Q^*(R) \equiv Q^*(R;b)$ can of course be obtained from the preceding formulas (5.11) to (5.15) inclusive for $Q(R) \equiv Q(R;b)$

[24] In this connection, Appendix D may be of interest.

by merely changing the integration limits correspondingly—for instance, in (5.11), from 0, R to R, ∞; in (5.13), from 0, L to L, ∞; and so on. However, the first four formulas for $Q^*(R)$ so obtained would suffer the disadvantage of each having an infinite limit of integration, rendering those formulas unsatisfactory for numerical integration purposes. This difficulty can be avoided by making the substitution $R = 1/r$ in each of those formulas for $Q^*(R)$. The resulting formulas are the following five, corresponding to (5.11) to (5.15) respectively:[24]

$$Q^*(R) = \frac{2}{\sqrt{1-b^2}} \int_0^r \frac{1}{\lambda^3} \exp\left[\frac{-1/\lambda^2}{1-b^2}\right] I_0\left[\frac{b/\lambda^2}{1-b^2}\right] d\lambda, \tag{5.16}$$

$$Q^*(R) = \frac{1}{\sqrt{1-b^2}} \int_0^{r^2} \frac{1}{\lambda^2} \exp\left[\frac{-1/\lambda}{1-b^2}\right] I_0\left[\frac{b/\lambda}{1-b^2}\right] d\lambda, \tag{5.17}$$

$$Q^*(R) = \sqrt{1-b^2} \int_0^{1/L} \frac{1}{\lambda^2} \exp\left[-\frac{1}{\lambda}\right] I_0\left[\frac{b}{\lambda}\right] d\lambda, \tag{5.18}$$

$$Q^*(R) = L\sqrt{1-b^2} \int_0^1 \frac{1}{\lambda^2} \exp\left[-\frac{L}{\lambda}\right] I_0\left[\frac{bL}{\lambda}\right] d\lambda, \tag{5.19}$$

$$Q^*(R) = \sqrt{1-b^2} \int_0^{\exp(-L)} I_0(b \log \lambda)\, d\lambda. \tag{5.20}$$

As a check on (5.16), it is obtainable from (4.2) by integrating the latter as to r.

For purposes of evaluation by numerical integration, formulas (5.11) to (5.15) inclusive may evidently differ greatly as regards the amount of labor involved and the numerical precision practically attainable. In each of these formulas except (5.14) the integrand contains only one parameter, b, while the integration range involves either R or $L \equiv R^2/(1-b^2)$. In (5.14) the integrand contains two independent parameters, b and L, while the integration range is a mere constant, 0-to-1. Similar statements apply to formulas (5.16) to (5.20) inclusive.

A partial check on any formula for $Q(R)$ can be applied by setting $R = \infty$, since $Q(\infty)$ should be equal to unity (representing certainty). If, for instance, this procedure is applied to formula (5.13), the right side is found to reduce to unity by aid of the known relation[25]

$$\int_0^\infty \exp(-A\lambda)\, J_0(B\lambda)\, d\lambda = \frac{1}{\sqrt{A^2+B^2}} \tag{5.21}$$

together with $I_0(B\lambda) = J_0(iB\lambda)$.

An integral formula of the secondary type for $Q^*(R) \equiv Q^*(R;b)$ can be obtained by substituting (2.20) into the last integral in (5.8), utilizing (2.25),

[25] Ref. 1, p. 384, Eq. (1); Ref. 2, p. 65, Eq. (2); Ref. 4, p. 58, Eq. (4.5).

changing the variable of integration by the substitution $\theta = \phi/2$, and rearranging; thus it is found that[26]

$$Q^*(R) = \frac{\sqrt{1-b^2}}{\pi \exp L} \int_0^\pi \frac{\exp(bL\cos\phi)}{1-b\cos\phi}\, d\phi. \tag{5.22}$$

This formula can also be obtained as a particular case of the more general formula (2.24) by setting $\theta = 2\pi$ in the upper limit of integration, utilizing (2.25), and changing the variable of integration by the substitution $\theta = \phi/2$.

Two partial checks on any general formula for $Q(R) \equiv Q(R;b)$ or for $Q^*(R) \equiv Q^*(R;b)$ can be applied by setting $b = 0$ and $b = 1$, and comparing the resulting particular formulas with those obtained by integrating the formulas for $P(R;0)$ and $P'(R;1)$ obtained in Section 3, namely formulas (3.20) and (3.21) there. It is thus found that

$$Q^*(R;0) = \exp(-R^2) = \int_R^\infty P(R;0)\, dR, \tag{5.23}$$

$$Q(R;1) = 2\left\{\frac{1}{\sqrt{2\pi}} \int_0^R \exp\left[-\frac{R^2}{2}\right] dR\right\} = \int_0^R P'(R;1)\, dR. \tag{5.24}$$

It will be recalled that the quantity between braces in (5.24) is extensively tabulated, and that it is sometimes called the 'normal probability integral.'

Several of the above general formulas for $Q(R) \equiv p(R'<R)$ and for $Q^*(R) \equiv p(R'>R)$ are closely connected with my 1933 paper.[27] Indeed, formulas (5.11), (5.14), (5.16), (5.19) and (5.22) above are the same as (53-A), (56-A), (52-A), (55-A) and (22-A), respectively, of the unpublished Appendix A to the 1933 paper; and (5.12), (5.13), (5.15), (5.17), (5.18) and (5.20) above were derived in the same connection, although they were not included in the Appendix A.

Formula (5.22) was employed in the unpublished Appendix A of the 1933 paper, being (22-A) there, as a basis for deriving two very different kinds of series type formulas for computing the values of $p(R'>R) \equiv Q^*(R)$ underlying the values of $p_{b,0}(R'>R)$ constituting Table I (facing Fig. 8) in that paper.[28]

[26] This formula, (5.22), was derived by me in a somewhat different manner in the unpublished Appendix A to my 1933 paper. Later I found that an equivalent formula, easily transformable into (5.22), had been given by Bravais as formula (51) in his classical paper "Analyse mathématique sur les probabilités des erreurs de situation d'un point," published in Mémoires de l'Académie Royale des Sciences de l'Institut de France, 2nd series, vol. IX, 1846, pp. 255–332. (This is available in the Public Library of New York City, for instance.)

[27] There the abbreviated symbols $p(R' < R)$ and $p(R' > R)$ were used with the same meanings as the complete symbols on the right sides of equations (5.1) and (5.2), respectively, of the present paper.

[28] Each of the two kinds of series type formulas comprised a finite portion of a convergent series plus an exact remainder term consisting of a definite integral. In the

In the present paper, formulas (5.11) and (5.16) have been used for numerical evaluation of $Q(R) \equiv p(R' < R)$ and of $Q^*(R) \equiv p(R' > R)$ by numerical integration (employing 'Simpson's one-third rule'), aided by some of the considerations set forth in Appendix D. However, only a moderate number of values of these quantities have been thus evaluated—merely enough to afford a fairly comprehensive check on Table I of my 1933 paper, by means of a sample consisting of 60 values (about 26%) distributed in a somewhat representative manner over that table. These new values of $Q^*(R) \equiv p(R' > R) = 1 - Q(R)$ are presented in Table 5.1 (at the end of this section) in such a way as to facilitate comparison with the old values, namely those in the 1933 paper. Thus, for any fixed value of R in Table 5.1, there are two horizontal rows of computed values of $Q^*(R)$, the first row (top row) coming from the 1933 paper, and the second row coming from the present paper. The third row of each set of four rows gives the deviations of the second row from the first row; and the fourth row expresses these deviations as percentages of the values in the first row.

In the first row of any set of four rows, any value represents $Q^*(R) \equiv p_b(R' > R)$ obtained, in accordance with Eq. (22) of my 1933 paper, by adding $\exp(-R^2)$ to $p_{b,0}(R' > R)$ given in Table I there. In the second row of a set, any value represents $Q^*(R) = 1 - Q(R)$ as computed by formula (5.11) or (5.16) of the present paper: more specifically, the values for $R = 0.2$, 0.4, 0.6 and 0.8 were computed by (5.11); and the values for $R = 1.6$ and $R = 2$ by (5.16), taking $r = 1/1.6 = 0.625$ and $r = 1/2 = 0.5$ respectively.[29]

In the 1933 paper, the values of $p_b(R' > R) \equiv Q^*(R;b)$ for $b = 0$ and for $b = 1$ were omitted as being unnecessary there because their values could be easily obtained from the simple exact formulas to which the general formulas there reduced, for $b = 0$ and $b = 1$. Those reduced formulas were the same as (5.23) and (5.24) of the present paper, except that (5.24) gives $Q(R;1)$ instead of giving $Q^*(R;1) = 1 - Q(R;1)$. The values obtained from these two formulas, exact to the number of significant figures here retained, are given in Table 5.1 at the intersections of the first row of each set of four rows with the columns $b = 0$ and $b = 1$. Therefore in these two columns the deviations (in the third row of each set of four rows) are deviations from exact values; the values in the second row of each set are, as

---

use of such a formula for numerical computations, the expansion producing the convergent series was carried far enough to insure that the remainder definite integral would be relatively small, though usually not negligible; and then this remainder definite integral was evaluated sufficiently accurately by numerical integration.

[29] In the work of numerical integration, 'Simpson's one-third rule' was employed for $R = 0.2$, 0.4, 0.6, 0.8 and 2. For $R = 1.6$, so that $r = 1/1.6 = 0.625$, 'Simpson's one-third rule' was employed up to $r = 0.620$, and the 'trapezoidal rule' from $r = 0.620$ to $r = 0.625$.

already stated, those obtained by the methods of the present paper, employing numerical integration.

From detailed inspection of Table 5.1 it will presumably be considered that the agreement between the two sets of values of $Q^*(R;b) \equiv p_b(R' > R)$ is to be regarded as satisfactory, at least from the practical viewpoint, the largest deviation being less than one per cent (for $R = 0.8$, $b = 0.9$).

TABLE 5.1
VALUES OF $Q^*(R) \equiv p(R' > R)$

b / R	0	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	1.00
0.2	.9608	.9590	.9574	.9550	.9516	.9463	.9372	.9168	.8930	.84148
"	.9623	.9605	.9590	.9567	.9528	.9473	.9387	.9206	.8925	.84124
"	.0015	.0015	.0016	.0017	.0012	.0010	.0015	.0038	−.0005	−.00024
"	.16	.16	.17	.18	.13	.11	.16	.41	−.06	−.03
0.4	.8521	.8462	.8410	.8335	.8228	.8071	.7830	.7420	.7127	.68916
"	.8537	.8477	.8427	.8351	.8240	.8081	.7841	.7459	.7125	.68897
"	.0016	.0015	.0017	.0016	.0012	.0010	.0011	.0039	−.0002	−.00019
"	.19	.18	.20	.19	.15	.12	.14	.53	−.03	−.03
0.6	.6977	.6880	.6799	.6686	.6531	.6324	.6055	.5721	.5578	.54851
"	.6992	.6892	.6814	.6698	.6540	.6334	.6065	.5764	.5572	.54831
"	.0015	.0012	.0015	.0012	.0009	.0010	.0010	.0043	−.0006	−.00020
"	.22	.17	.22	.18	.14	.16	.17	.75	−.11	−.04
0.8	.5273	.5167	.5081	.4969	.4826	.4656	.4477	.4316	.4261	.42371
"	.5290	.5183	.5099	.4982	.4840	.4672	.4488	.4357	.4266	.42355
"	.0017	.0016	.0018	.0013	.0014	.0016	.0011	.0041	.0005	−.00016
"	.32	.31	.35	.26	.29	.34	.25	.95	.12	−.04
1.6	.07730	.07986	.08207	.08522	.0891	.0938	.0990	.1042	.1070	.10960
"	.07727	.07988	.08210	.08536	.0892	.0938	.0989	.1042	.1069	.10958
"	−.00003	.00002	.00003	.00014	.0001	.0000	−.0001	.0000	−.0001	−.00002
"	−.04	.03	.04	.16	.11	.00	−.10	.00	−.09	−.02
2.0	.01832	.02153	.02394	.02681	.0301	.0337	.0375	.0414	.0435	.04550
"	.01823	.02145	.02383	.02685	.0302	.0338	.0376	.0415	.0436	.04552
"	−.00009	−.00008	−.00011	.00004	.0001	.0001	.0001	.0001	.0001	.00002
"	−.49	−.37	−.46	.15	.33	.30	.27	.24	.23	.04

6. THE DISTRIBUTION FUNCTION FOR THE ANGLE

The distribution function $P(\theta \mid R_{12}) \equiv P(\theta)$ for the angle θ of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.10) on setting $\rho = \theta$, $\sigma = R$, $\sigma_1 = R_1 = 0$ and $\sigma_2 = R_2 = \infty$; thus

$$P(\theta)d\theta = p(\theta < \theta' < \theta + d\theta,\ 0 < R' < \infty). \tag{6.1}$$

An integral formula for $P(\theta)$ is immediately obtainable from (1.6) by setting $\rho = \theta$, $\sigma = R$, $\sigma_3 = \sigma_1 = R_1 = 0$ and $\sigma_4 = \sigma_2 = R_2 = \infty$; thus

$$P(\theta) = \int_0^\infty P(R, \theta)\, dR. \tag{6.2}$$

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $P(\theta)$ is here an abbreviation for $P(\theta;b)$.

A formula for $P(\theta;b) \equiv P(\theta)$ can be obtained by substituting $P(R,\theta)$ from (2.15) into (6.2) and executing the indicated integration, which can be easily accomplished. The resulting formula is found to be

$$P(\theta;\, b) = \frac{\sqrt{1-b^2}}{2\pi(1-b\cos 2\theta)}. \tag{6.3}$$

This formula can also be obtained as a particular case of either of the more general formulas (2.19) and (2.20) by setting $R = \infty$ in (2.19) or $R = 0$ in (2.20); also by adding (2.19) to (2.20) and then utilizing (1.10).

In $P(\theta) \equiv P(\theta;b)$ it will evidently suffice to deal with values of θ in the first quadrant, because of symmetry of the scatter diagram.

In $P(\theta;b)$ it will suffice to deal with only positive values of b, as (6.3) shows that changing b to $-b$ has the same effect as changing 2θ to $\pi\pm 2\theta$, or θ to $\pi/2\pm\theta$; that is, $P(\theta;-b) = P(\pi/2\pm\theta;b)$.

Fig. 6.1 gives curves of $P(\theta;b)$, computed from (6.3), as function of θ with b as parameter, for the ranges[30] $0\leqq\theta\leqq 90°$ and $0\leqq b\leqq 1$.

The curves in Fig. 6.1 indicate that $P(\theta;b)$ is a maximum at $\theta = 0°$ and a minimum at $\theta = 90°$. These indications are verified by formula (6.3), as this formula shows that:

$$\text{Max } P(\theta;b) = P(0°;b) = \frac{1}{2\pi}\sqrt{\frac{1+b}{1-b}}, \tag{6.4}$$

$$\text{Min } P(\theta;b) = P(90°;b) = \frac{1}{2\pi}\sqrt{\frac{1-b}{1+b}}. \tag{6.5}$$

Thence

$$\text{Min } P(\theta;b)/\text{Max } P(\theta;b) = (1-b)/(1+b), \tag{6.6}$$

$$P(\theta;b)/\text{Max } P(\theta;b) = P(\theta;b)/P(0°;b) = (1-b)/(1-b\cos 2\theta). \tag{6.7}$$

The curves in Fig. 6.1 indicate also that $P(\theta;b)$ is independent of θ in the case $b = 0$. This is verified by formula (6.3), as this formula shows that

$$P(\theta;0) = 1/2\pi. \tag{6.8}$$

Thence (6.3) can be written

$$P(\theta;b)/P(\theta;0) = (\sqrt{1-b^2})/(1-b\cos 2\theta). \tag{6.9}$$

[30] Beginning here, θ will usually be expressed in degrees instead of radians, for practical convenience.

By setting $\cos 2\theta = 0$ in (6.3), so that $\theta = 45°$, it is found that

$$(\sqrt{1 - b^2})/2\pi = P(45°;b), \tag{6.10}$$

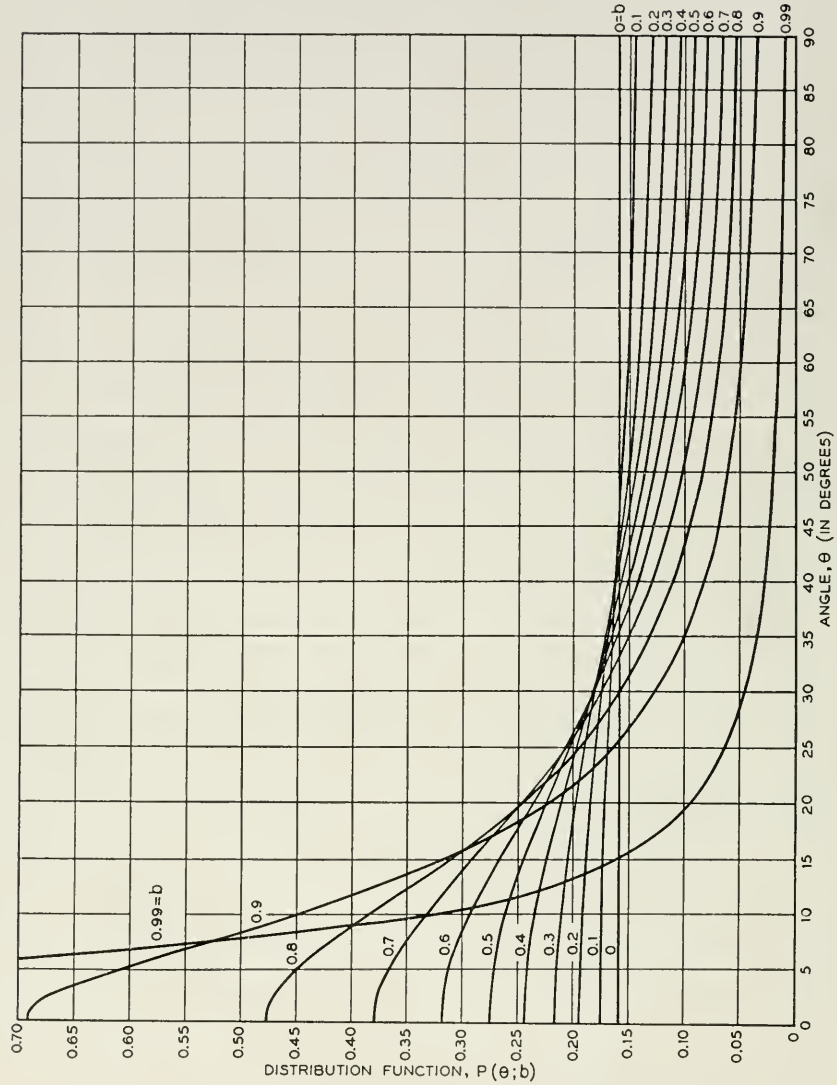


Fig. 6.1—Distribution function for the angle.

whence (6.3) can be written

$$P(\theta;b)/P(45°;b) = 1/(1-b\cos 2\theta). \tag{6.11}$$

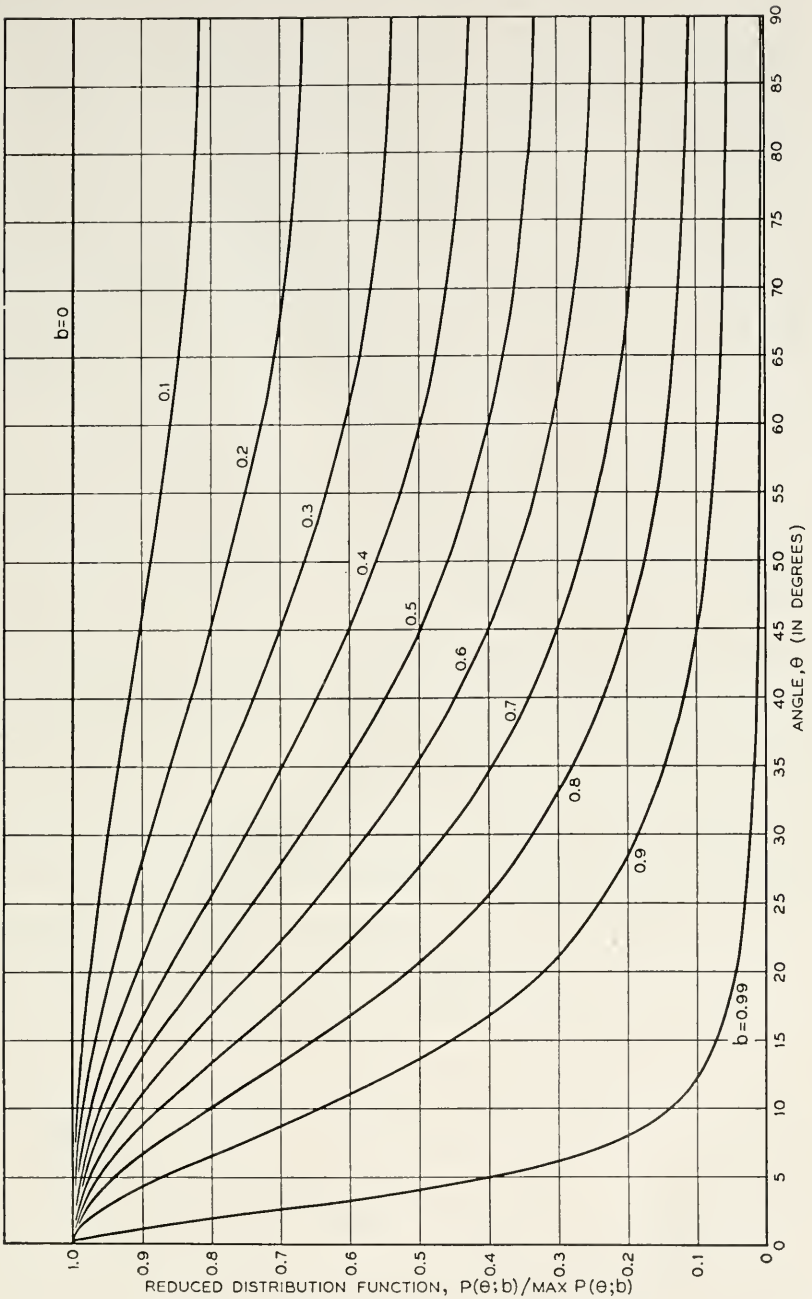


Fig. 6.2—Reduced distribution function for the angle.

In the case $b = 1$, the curves in Fig. 6.1 suggest, by limiting considerations, that $P(\theta;1)$ is zero for all θ except $\theta = 0°$, and that $P(\theta;1)$ is infinite for $\theta = 0°$. These conclusions are verified by formula (6.3), as this formula shows that:

$$P(\theta;1) = 0 \text{ for } 0° < \theta < 180°;\ P(\theta;1) = \infty \text{ for } \theta = 0°,\ 180°.$$

The curves in Fig. 6.1, though having the advantage of directly representing $P(\theta;b)$ as function of θ with b as parameter, are somewhat troublesome to use because of their numerous crossings of each other. This difficulty is not present in Fig. 6.2, which gives curves of $P(\theta;b)/\text{Max}\ P(\theta;b)$, obtained by dividing the ordinates $P(\theta;b)$ of the curves in Fig. 6.1 by the respective maximum ordinates of those curves, as given by (6.4), so that the equation of the curves in Fig. 6.2 is formula (6.7).

7. The Cumulative Distribution Function for the Angle

The cumulative distribution function $Q(<\theta,R_{12}) \equiv Q(\theta)$ for the angle θ of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.11) on setting $\rho = \theta$, $\sigma = R$, $\rho_1 = \theta_1 = 0$, $\sigma_1 = R_1 = 0$ and $\sigma_2 = R_2 = \infty$; thus

$$Q(\theta) = p(0 < \theta' < \theta,\ 0 < R' < \infty). \tag{7.1}$$

A 'double integral' for $Q(\theta)$, in the form of two 'repeated integrals,' can be written down directly by inspection of the $p(\ )$ expression in (7.1) or by specialization of (1.8); thus

$$Q(\theta) = \int_0^\theta \left[\int_0^\infty P(R,\theta)\,dR\right] d\theta = \int_0^\infty \left[\int_0^\theta P(R,\theta)\,d\theta\right] dR. \tag{7.2}$$

Evidently these can be written formally as two 'single integrals,'

$$Q(\theta) = \int_0^\theta P(\theta)\,d\theta = \int_0^\infty P(R \mid <\theta)\,dR, \tag{7.3}$$

by means of the distribution functions $P(\theta) \equiv P(\theta \mid R_{12})$ and $P(R \mid <\theta)$ given by the formulas

$$P(\theta) = \int_0^\infty P(R,\theta)\,dR, \tag{7.4}$$

$$P(R \mid <\theta) = \int_0^\theta P(R,\theta)\,d\theta. \tag{7.5}$$

(7.4) is the same as (6.2). (7.5) is a special case of (1.6), and the left side of (7.5) is a special case of $P(\rho \mid <\sigma)$ defined by (1.13).

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $Q(\theta)$ is here an abbreviation for $Q(\theta;b)$.

A natural and convenient way for deriving formulas for $Q(\theta)$ is afforded

by the general formula (7.3) together with the auxiliary general formulas (7.4) and (7.5), beginning with the two latter.

It will be convenient to dispose of (7.5) before dealing with (7.4), as (7.5) turns out to be the less useful. For when $P(R,\theta)$ given by (2.16) is substituted into (7.5), the indicated integration cannot be executed in general, as (7.5) becomes (2.18), wherin the indicated integration can be executed only for certain special values of the integration limit θ—by means of the special Bessel function formula (3.3).

When $P(R,\theta)$ given by (2.15), which is equivalent to (2.16) used above, is substituted into (7.4), it is found that the indicated integration can be executed, giving the previously obtained formula (6.3) for $P(\theta) \equiv P(\theta;b)$.

A θ-integral formula for $Q(\theta) \equiv Q(\theta;b)$ can be obtained by substituting $P(\theta) \equiv P(\theta;b)$ from (6.3) into the first integral in (7.3), giving

$$Q(\theta;\, b) = \frac{\sqrt{1-b^2}}{2\pi}\int_0^{\theta}\frac{d\theta}{1-b\cos 2\theta} = \frac{\sqrt{1-b^2}}{4\pi}\int_0^{2\theta}\frac{d\phi}{1-b\cos\phi}. \quad (7.6)$$

This formula can also be obtained as a particular case of the more general formulas (2.22) and (2.24) by setting $R = \infty$ in (2.22) or $R = 0$ in (2.24); also by adding (2.22) to (2.24) and then utilizing (1.11).

The integral in (7.6) is of well-known form, and the indicated integration can be executed, yielding the following two equivalent formulas for $Q(\theta;b)$:

$$\begin{aligned} Q(\theta;\, b) &= \frac{1}{2\pi}\left|\tan^{-1}\left[\sqrt{\frac{1+b}{1-b}}\tan\theta\right]\right| \\ &= \frac{1}{4\pi}\left|\cos^{-1}\left[\frac{\cos 2\theta - b}{1-b\cos 2\theta}\right]\right|. \end{aligned} \quad (7.7)$$

In $Q(\theta;b)$ it will evidently suffice to deal with values of θ in the first quadrant, because of symmetry of the scatter diagram, and the resulting fact that $Q(n\,90°) = n/4$, where $n = 1, 2, 3$ or 4.

In $Q(\theta;b)$ it will suffice to deal with positive values of b, as (7.7) shows that[31]

$$Q(\theta;\, -b) = \left|\frac{1}{4} - Q\left(\frac{\pi}{2} \pm \theta;\, b\right)\right|.$$

Fig. 7.1 gives curves of $Q(\theta;b) \equiv Q(\theta)$ computed from (7.7), as function of θ with b as parameter, for the ranges $0 \leqq \theta \leqq 90°$ and $0 \leqq b \leqq 1$.

Consideration of the scatter diagram of W or of its equiprobability curves, which are concentric similar ellipses, affords several partial checks on the curves in Fig. 7.1 and on formula (7.7) from which they were plotted.

[31] This relation can also be derived geometrically from the fact that the scatter diagram for $-b$ is obtainable by merely rotating that for b through 90°, as shown by (2.6), or (2.7) and (2.8), or (2.11).

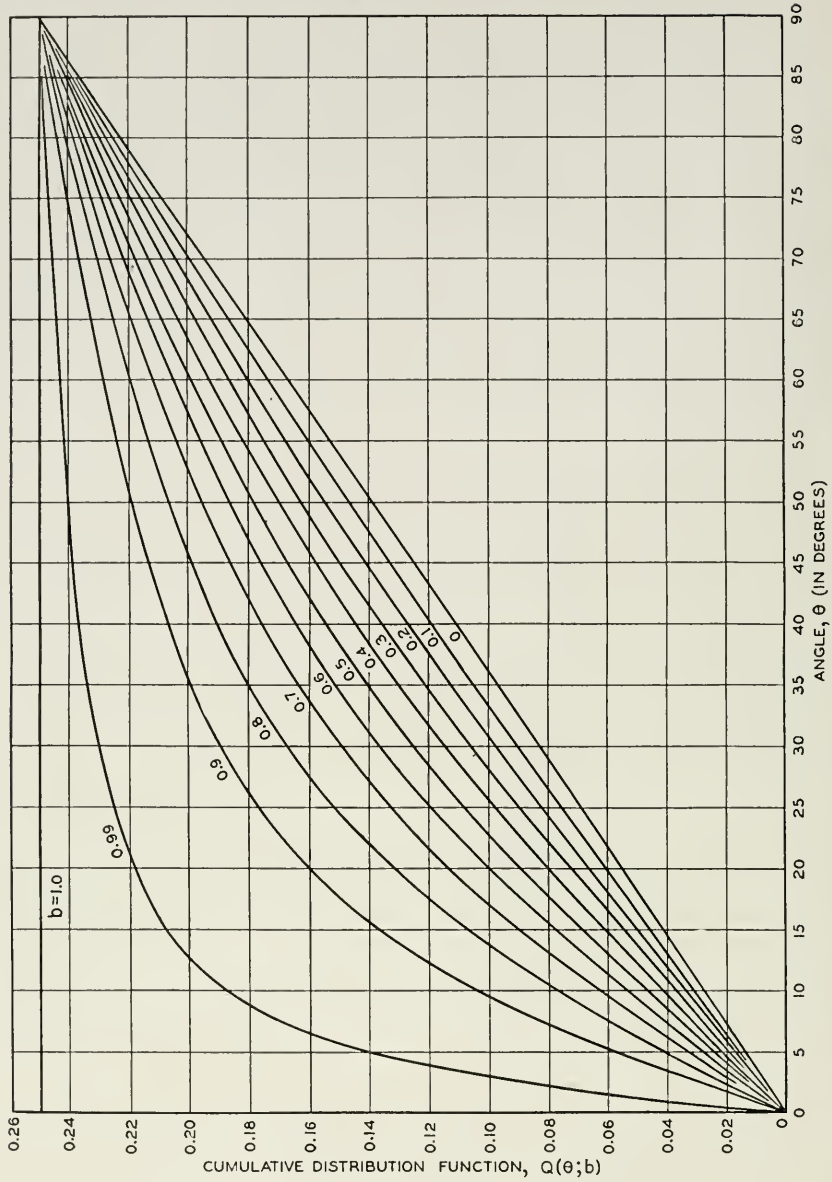


Fig. 7.1—Cumulative distribution function for the angle.

Thus, the fact that the curve for $b = 0$ is a straight line, whose equation is

$$Q(\theta;0) = \theta/2\pi = \theta^\circ/360^\circ, \qquad (b = 0),$$

corresponds to the fact that for $b = 0$ the equiprobability curves are circles.

The fact that the curve for $b = 1$ is the straight line $Q(\theta;1) = 1/4 = 0.25$ corresponds to the fact that for $b = 1$ the scatter diagram has degenerated to be merely a straight line coinciding with the real axis, so that no point outside of this line makes any contribution to $Q(\theta;1)$.

The fact that, at $\theta = 90°$, $Q(\theta;b) = Q(90°;b)$ has for all b the value $1/4 = 0.25$ corresponds to the fact that the area of a quadrant of the scatter diagram is one-fourth the area of the entire scatter diagram. Hence $Q(360°;b) = 4Q(90°;b) = 1$, which is evidently correct.

Acknowledgment

The computations and curve-plotting for this paper were done by Miss M. Darville; those for the 1933 paper, by Miss D. T. Angell.

APPENDIX A

Derivation of Formula (2.15) for $P(R,\theta)$

(2.15) will here be derived from (2.11) by utilizing the fact that the 'areal probability density', G, at any fixed point in the scatter diagram must be independent of the system of coordinates; for $G\,dA$ gives the probability of falling in any differential element of area dA, and this probability must evidently be independent of the shape of dA (assuming that all linear dimensions of dA are differential, of course). Thus, indicating the element of area by an underline, we have, in rectangular coordinates,

$$\underline{GdUdV} = P(U,V)\underline{dUdV}, \qquad \text{(A1)} \qquad \text{whence} \quad G = P(U,V). \qquad \text{(A2)}$$

In polar coordinates,

$$\underline{GRd\theta dR} = P(R,\theta)\underline{dRd\theta}, \qquad \text{(A3)} \qquad \text{whence} \quad G = P(R,\theta)/R. \qquad \text{(A4)}$$

Comparing these two expressions for G shows that[32]

$$P(R,\theta) = RP(U,V). \qquad \text{(A5)}$$

Thus, a formula for $P(R,\theta)$ can be obtained from (2.11) by merely multiplying both sides of that formula by R. However, in the resulting formula it will remain to express U and V in terms of R and θ, by means of the relations

$$U = R\cos\theta, \qquad \text{(A6)} \qquad V = R\sin\theta. \qquad \text{(A7)}$$

The final result, after a simple reduction, is (2.15), which is thus proved.

APPENDIX B

Formulas of the Curves in Fig. 3.3

As in equation (3.22), R_c will here denote the critical value of R, that is, the value of R at which $P(R) \equiv P(R;b)$ has its maximum value; and T_c

[32] Formula (A5) can be easily verified by the entirely different method which utilizes (1.23).

will denote the corresponding value of T, whence T_c is given in terms of R_c and b by (3.22).

A formula for $dP(R)/dR$ could of course be obtained directly from (3.4) but it will be found preferable to obtain it indirectly from the less cumbersome formula (3.8) containing the auxiliary variable T defined by (3.6). Evidently, since b does not depend on R,

$$\frac{dP(R)}{dR} = \frac{dP(R)}{dT}\frac{dT}{dR} = \frac{2bR}{1-b^2}\frac{dP(R)}{dT}. \tag{B1}$$

Thus, since the factor $2bR/(1-b^2)$ cannot vanish for any value of R (except $R = 0$), the only critical value of R must be that corresponding to the value of T at which $dP(R)/dT$ vanishes, namely T_c, since T_c has been defined to be the value of T corresponding to R_c. (Incidentally, equation (B1) shows that T_c is equal to the value of T at which $P(R)$ is an extremum when $P(R)$ is regarded as a function of T.) From (3.22),

$$\frac{R_c^2}{1-b^2} = \frac{T_c}{b}. \tag{B2}$$

Evidently T_c and R_c must ultimately be functions of only b. The next paragraph deals with T_c, which evidently has to be known before R_c can be evaluated.

From (3.8) it is found that, since $dI_0(T)/dT = I_1(T)$,

$$\frac{dP(R)}{dT} = P(R)\left[\frac{1}{2T} + \frac{I_1(T)}{I_0(T)} - \frac{1}{b}\right]. \tag{B3}$$

Hence, since $P(R)$ does not vanish for any value of R (except $R = 0$ and $R = \infty$), T_c will be a root of the conditional equation obtained by equating to zero the expression in brackets in (B3). This conditional equation is transcendental in T_c and apparently has no closed form of explicit solution for T_c; and its solution by successive approximation, or otherwise, would likely be rather slow and laborious. However, the bracket expression in (B3) shows that b can be immediately expressed explicitly in terms of T_c by the equation

$$b = \frac{2T_c}{1 + 2T_c I_1(T_c)/I_0(T_c)}. \tag{B4}$$

For some purposes, the following two equations, each equivalent to (B4), will be found more convenient:

$$\frac{T_c}{b} = \frac{1}{2} + T_c\frac{I_1(T_c)}{I_0(T_c)}, \tag{B5}$$

$$\frac{T_c}{b} = \frac{1/2}{1 - bI_1(T_c)/I_0(T_c)}. \tag{B6}$$

On account of (B2), the right sides of (B5) and (B6) are equal not only to T_C/b but also to $R_C^2/(1-b^2)$.

Since the utilization of formulas (B4), (B5) and (B6) for computing the curves in Fig. 3.3 will involve taking T_C as the independent variable and assigning to it a set of chosen numerical values, the natural first step is to find approximately the range of T_C corresponding to the b-range, $0 \leqq b \leqq 1$, in order to be able to choose only useful values of T_C. This step will be taken in the next paragraph.

Equation (B6) shows that $T_C/b = 1/2$ when $b = 0$, and hence that $T_C = 0$ when $b = 0$; and this last is verified by (B4). The other end-value of the T_C-range, namely the value of T_C for $b = 1$, cannot be found explicitly and exactly. However, rough values of limits between which it must lie can be found fairly easily as follows: To begin with, each of the equations (B5) and (B6) shows that $T_C \geqq b/2$, for all values of b in $0 \leqq b \leqq 1$; in particular, $T_C > 1/2$ when $b = 1$. An upper limit for T_C for any value of b can be found from (B5) by utilizing the power series expressions for $I_1(T_C)$ and $I_0(T_C)$, whereby it is found that

$$\frac{I_1(T_C)}{I_0(T_C)} = H\frac{T_C}{2}, \qquad \text{(B7)} \qquad \text{where} \qquad H \approx 1 - \frac{T_C^2}{8} < 1. \qquad \text{(B8)}$$

On substituting (B7) into (B5) and then solving for T_C in terms of b and H, it is found that

$$T_C = b/(1 + \sqrt{1 - Hb^2}). \qquad \text{(B9)}$$

On account of (B8), (B9) shows that

$$T_C < b/(1 + \sqrt{1 - b^2}), \qquad \text{(B10)}$$

whence, in particular, $T_C < 1$ when $b = 1$. By successive approximation or otherwise, it can now be rather quickly found that, when $b = 1$, $T_C = 0.79$ (to two significant figures).[33]

From the preceding paragraph, it is seen that, when b ranges from 0 to 1, T_C ranges from 0 to about 0.79; T_C/b ranges from 0.5 to about 0.79; and, on account of (B2), R_C ranges from $\sqrt{0.5} = 0.707$ down to 0.

The curves in Fig. 3.3 are constructed with the aid of the formulas and methods of this appendix as follows: First, a set of values of T_C is chosen, ranging from 0 to 0.79 and slightly larger. Second, for each such chosen T_C the right side of (B5) is computed, thereby evaluating T_C/b and also $R_C^2/(1-b^2)$, these two quantities being equal by (B2). Third, the corresponding value of b is found by dividing T_C by T_C/b; less easily, it could

[33] Because of the special importance of $b = 1$ in other connections, T_C for $b = 1$ was later evaluated to four significant figures and found to be $T_C = 0.7900$; thence, by substituting this value of T into (3.8), along with $b = 1$, it was found that Max. $P(R;1) = 0.9376$, which occurs at $R = R_C = 0$, by (B2).

be found by substituting T_c into (B4). Fourth, from T_c/b the value of $\sqrt{T_c/b}$ is found, and thereby the value of $R_c/\sqrt{1-b^2}$ and thence R_c. Finally, Max. $P(R;b)$ is computed by inserting the critical values into any of the various (equivalent) formulas for $P(R;b)$, namely (3.4), (3.7), (3.8), (3.10) or (3.12).

APPENDIX C

Fomulas of the Curves in Fig. 4.3

The first six equations of this appendix are given without derivation and almost without any comments because they correspond exactly and simply to the first six equations, respectively, of Appendix B. Beginning with the second paragraph of the present appendix, the close correspondence ceases.

$$\frac{dP(r)}{dr} = \frac{dP(r)}{dT}\frac{dT}{dr} = \frac{-2b}{(1-b^2)r^3}\frac{dP(r)}{dT}. \tag{C1}$$

$$\frac{1}{(1-b^2)r_c^2} = \frac{T_c}{b}. \tag{C2}$$

$$\frac{dP(r)}{dT} = P(r)\left[\frac{3}{2T} + \frac{I_1(T)}{I_0(T)} - \frac{1}{b}\right]. \tag{C3}$$

$$b = \frac{2T_c}{3 + 2T_c\, I_1(T_c)/I_0(T_c)}. \tag{C4}$$

$$\frac{T_c}{b} = \frac{3}{2} + T_c\,\frac{I_1(T_c)}{I_0(T_c)}. \tag{C5}$$

$$\frac{T_c}{b} = \frac{3/2}{1 - bI_1(T_c)/I_0(T_c)}. \tag{C6}$$

The bracketed expression in (C3) is seen to be obtainable from that in (B3) by merely changing T to $T/3$ wherever T does not occur as the argument of a function; hence the three equations following (C3) are obtainable from the three equations following (B3) by correspondingly changing T_c to $T_c/3$. (In this appendix, as in Section 4, small c is purposely used as a subscript to indicate a 'critical' value, whereas in Section 3 and in Appendix B, capital C is used for that purpose.)

For use below, it will here be noted that

$$I_1(T_c)/I_0(T_c) = N_1(T_c)/N_0(T_c), \tag{C7}$$

as will be seen by dividing (3.16) by (3.13). On account of (3.17) and (3.14), (C7) shows that for large values of T_c the right side of (C7) is equal to 1 as a first approximation, and to $1 - 1/2T_c$ as a second approximation; thus, for large T_c,

$$I_1(T_c)/I_0(T_c) \approx 1 - 1/2T_c \approx 1. \qquad \text{(C8)}$$

The first step toward computing the curves in Fig. 4.3 is to find approximately the T_c-range corresponding to the b-range, $0 \leqq b \leqq 1$. This is done in the course of the next four paragraphs.

When $b = 0$, equation (C6) shows that $T_c/b = 3/2$ and hence that $T_c = 0$; or, what is equivalent, $b/T_c = 2/3$ and hence $1/T_c = \infty$ (since $b = 0$).

When $b = 1$, $T_c = \infty$, as can be easily verified from equation (C4), (C5) or (C6) by utilizing (C8).

Thus, from the two preceding paragraphs, it is seen that, when b ranges from 0 to 1, b/T_c ranges from 2/3 to 0; T_c/b from 3/2 to ∞; and T_c from 0 to ∞.

Since $T_c = \infty$ when $b = 1$, the choosing of a set of finite values of T_c will necessitate an approximate formula for computing T_c for values of b nearly equal to 1, which means for very large values of T. Such a formula is easily obtainable from (C5) by utilizing the approximation $1 - 1/2T_c$ in (C8), whereby it is found that, for large T_c,

$$T_c \approx b/(1-b), \qquad \text{(C9)} \qquad\qquad b/T_c \approx 1-b. \qquad \text{(C10)}$$

As examples, these approximate formulas give: When $b = 0.99$, $T_c \approx 99$, $b/T_c \approx 0.01$; when $b = 0.9$, $T_c \approx 9$, $b/T_c \approx 0.1$. It will be found that even in the second example the results are pretty good approximations.

The curves in Fig. 4.3 are constructed with the aid of the formulas and methods of this appendix as follows: First, a set of values of T_c is chosen, ranging from 0 to about 100 (the latter figure corresponding approximately to $b = 0.99$). Second, for each such chosen T_c the right side of (C5) is computed, thereby evaluating T_c/b and also $1/(1-b^2)r_c^2$, these two quantities being equal by (C2). Third, the corresponding value of b is found by dividing T_c by T_c/b; less easily, it could be found by substituting T_c into (C4). Fourth, from T_c/b the value of $\sqrt{T_c/b}$ is found, and thereby the value of $1/r_c\sqrt{1-b^2}$ and thence r_c. Finally, Max $P(r;b)$ is computed by inserting the critical values into any of the (equivalent) formulas for $P(r;b)$, namely (4.2), (4.3) or (4.4).

APPENDIX D

Some Simple General Considerations Regarding the Evaluation of Cumulative Distribution Functions by Numerical Integration

This appendix gives some simple general considerations and relations that may sometimes facilitate and render more accurate the evaluation of cumulative distribution functions by numerical integration.

Some of these considerations and relations have found application in Section 5 in the evaluation of the cumulative distribution function for the modulus $R \equiv |W|$. For this reason, the variate in the present section will be denoted by R, though without thereby restricting R to denote the modulus; rather, R will here denote any positive real variate, though it should preferably be a 'reduced' variate, so as to be dimensionless, as in equation (2.9). The restriction of R to positive values is imposed because it is strongly conducive to simplicity and brevity of treatment, without constituting an ultimate limitation. The reciprocal of R will be denoted by r, as previously.[34]

We may wish to evaluate numerically the cumulative distribution function $p(R'<R) \equiv Q(R)$ or $p(R'>R) \equiv Q^*(R)$ or both. Since these are not independent, their sum being equal to unity, the evaluation of either one determines the other, theoretically. However, when the evaluated one is nearly equal to unity, the remaining one may perhaps not be evaluable with sufficient accuracy (percentagewise) by subtracting the evaluated one from unity. Then it would presumably be advantageous to introduce for auxiliary purposes the variable $r = 1/R$, since evidently

$$p(R'>R) = p(1/R'<1/R) = p(r'<r), \tag{D1}$$

$$p(R'<R) = p(r'>r) = 1 - p(r'<r). \tag{D2}$$

Thus, if $p(R'>R)$, in (D1), is small compared to unity, it is presumably evaluable with higher accuracy percentagewise by dealing with $p(r'<r)$ than with $1 - p(R'<R)$. Incidentally, after $p(r'<r)$ has been evaluated, it might be used in (D2) to arrive at a still more accurate value of $p(R'<R)$ than had originally been obtained directly by numerical integration.

Assuming that we have a plot (or a table) of the distribution function $P(R)$, we can evidently evaluate

$$P(R'<R^0) = \int_0^{R^0} P(R)dR \tag{D3}$$

directly by numerical integration, provided the plot is sufficiently extensive to include R^0; if not, we can, by (D2), resort to

$$p(R'<R^0) = 1 - p(r'<r^0) = 1 - \int_0^{r^0} P(r)dr, \tag{D4}$$

assuming that a sufficiently extensive plot (or table) of $P(r)$ is available and applying numerical integration to it.

Even if the plot of $P(R)$ used in (D3) is sufficiently extensive to include

[34] The restriction of R, and hence of r, to positive values is seen to be absent from equations (D1), (D2), (D5) and (D6) but present in (D3), (D4), (D7) and (D8).

R^0, so that (D3) could be evaluated, it might be that (D4) would result in greater accuracy; this would presumably be the case when $p(R'<R^0)$ is nearly equal to unity.

Evidently an evaluation of

$$p(R'>R^0) = \int_{R^0}^{\infty} P(R)dR \tag{D5}$$

directly by numerical integration would be less satisfactory than the evaluation of $p(R'<R^0)$ in the preceding paragraph. For, due to the presence of the infinite limit in the integral in (D5), the plot of $P(R)$ would have to be carried to a large enough value of R so that the integral from there to ∞ would be known to be negligible. This difficulty can be avoided by starting with the relation

$$p(R'>R^0) = 1 - p(R'<R^0) \tag{D6}$$

and substituting therein the value of $p(R'<R^0)$ given by (D3) or (D4), resulting respectively in the following two formulas:

$$p(R'>R^0) = 1 - \int_0^{R^0} P(R)dR, \tag{D7}$$

$$p(R'>R^0) = p(r'<r^0) = \int_0^{r^0} P(r)dr, \tag{D8}$$

the integrals in which are evidently suitable for evaluation by numerical integration, none of the integration limits being infinite. If $p(R'>R^0)$ is small compared to unity, (D8) would presumably be more accurate (percentagewise) than (D7). If the plot of $P(R)$ is not sufficiently extensive to include R^0, (D7) evidently could not be used; but, instead, (D8) could be used if the plot of $P(r)$ were sufficiently extensive to include r^0.

References on Bessel Functions

1. Watson, "Theory of Bessel Functions," 1st. Ed., 1922; or 2nd Ed., 1944.
2. Gray, Mathews and MacRobert, "Bessel Functions," 2nd Ed., 1922.
3. McLachlan, "Bessel Functions for Engineers," 1934.
4. Bowman, "Introduction to Bessel Functions," 1938.
5. Whittaker and Watson, "Modern Analysis," 2nd Ed., 1915.
6. "British Association Mathematical Tables," Vol. VI: Bessel Functions, Part I, 1937.
7. Anding, "Sechsstellige Tafeln der Bessel'schen Funktionen imaginaren Arguments," 1911 (mentioned on p. 657 of Ref. 1).

Spectrum Analysis of Pulse Modulated Waves

By J. C. LOZIER

The problem here is to find the frequency spectrum produced by the simultaneous application of a number of frequencies to various forms of amplitude limiters or switches. The method of solution presented here is to first resolve the output wave into a series of rectangular waves or pulses and then to combine the spectrum of the individual pulses by vectorial means to find the spectrum of the output. The rectangular wave shape was chosen here as the basic unit in order to make the method easy to apply to pulse modulators.

Introduction

The rapidly expanding use of pulse modulation[1] in its various forms is bound to make the frequency spectrum of pulse modulated waves a subject of increasing practical importance. The purpose of this paper is to show how to determine the frequency spectrum of these waves by methods based as far as possible on physical rather than mathematical considerations. The physical approach is used in an attempt to maintain throughout the analysis a picture of the way in which the various factors contribute to a given result. To further this objective the fundamentals involved are reviewed from the same point of view.

The method is used here to analyze two distinct types of pulse modulation, namely, pulse position and pulse width modulation.[2] These two cases are especially important for illustrative purposes because their spectra can be tied back to more familiar methods of modulation. Thus it will be shown that, as the ratio of the pulse rate to the signal frequency becomes large, pulse position modulation becomes a phase modulation of the various carrier frequencies that form the frequency spectrum of the unmodulated pulse wave, and pulse width modulation becomes a form of amplitude modulation of its equivalent carriers. The analysis also shows certain interesting input-output relationships that may be obtained from such modulators, treating them as straight transmission elements at the signal frequency.

These relationships are of more than theoretical interest. The pulse position modulator has already been used as phase or frequency modulator to good advantage.[3] The use of a pulse width modulator as an amplifier is

[1] E. M. Deloraine and E. Labin, "Pulse Time Modulation", *Electrical Communications*, Vol. 22, No. 2, pp. 91-98, Dec. 1944; H. S. Black "AN-TRC-6 A Microwave Relay System", *Bell Labs. Record*, V. 33, pp. 445–463, Dec. 1945.

[2] By *pulse position* modulation is meant that form of pulse modulation in which the length of each pulse is kept fixed but its position in time is shifted by the modulation, and by *pulse width* modulation that form in which the length of each pulse varies with the modulation but the center of each pulse is not shifted in position.

[3] L. R. Wrathall, "Frequency Modulation by Non-linear Coils", *Bell Labs. Record*, Vol. 23, pp. 445–463, Dec. 1945.

another practical application, of which the self oscillating or hunting servomechanism is an example.

The quantitative analysis of such systems depends on the ratio of the pulse repetition rate to the signal frequency. When this ratio is low, the solution can be obtained by a method shown here for resolving the modulated waves into selected groups of effectively unmodulated components. This technique is powerful since it can be done by graphical means whenever the complexity of either the system or the signal warrants it. When the ratio of pulse rate to signal frequency becomes high enough, such methods are no longer practical. However, under these conditions other methods become available, especially in cases like those mentioned above where the spectrum of the modulation approaches one of the more familiar forms. An important example of this occurs in the case of the pulse position modulator where, as the spectrum approaches that of phase modulated waves, the solution can often be found by the conventional Bessel's function technique used in analyzing phase and frequency modulators.

The method proposed here for obtaining the spectrum analysis of pulse modulated waves is based on the use of the magnitude-time characteristic of the single pulse and its frequency spectrum as a pair of interchangeable building blocks, so that the analysis will develop this relationship. Before doing this the elementary theory of spectrum analysis will be reviewed

Review of the Elementary Theory of Spectrum Analysis

A complex wave may be represented in two ways. One way is by its magnitude at each instant of time. The other way is by its frequency spectrum, that is, by the various sinusoidal components that go to make up the wave. The two representations are interchangeable.

The transformation from a given frequency spectrum to the corresponding magnitude vs. time function is straight-forward, for it is apparent that the various components in the frequency spectrum must add up to the desired magnitude-time function. The necessary additions may be difficult to make in some cases but they are not hard to understand.

The reverse process of finding the frequency spectrum when the magnitude-time characteristic is given is more involved, though using Fourier analysis, the problem can generally be formulated readily enough. Furthermore the mathematical procedures involved can be interpreted physically in broad terms by modulation theory. However, these procedures become more difficult to perform, and the physical relationships more obscure, as the wave form under analysis becomes more complex. This is particularly true when general or informative solutions rather than specific answers are required. Pulse modulated waves are sufficiently new and complex to give such difficulties.

The process of finding the frequency spectrum of a complex wave from its magnitude-time function has a simple mathematical basis. It depends on the fact that the square of a sinusoidal wave has a positive average value over any interval of time, whereas the product of two sinusoidal waves of different frequencies will average zero over a properly chosen interval of time.[4]

In theory then, as the magnitude-time function of a complex wave is the sum of all the components of the frequency spectrum, we have only to multiply this magnitude-time function by a sinusoidal wave of the desired frequency and then average the product over the proper time interval to find the component of the spectrum at this frequency.[5]

One physical interpretation of this procedure can be given in terms of modulation theory. The product of the magnitude-time function with a sinusoidal wave will produce the beat or sum and difference frequencies between the frequency of the sinusoid and each component of the frequency spectrum. Thus, if the spectrum contains the same frequency, a zero beat or dc term is produced, and this term may be evaluated by averaging the product over an interval that is of the proper length to make all the ac components vanish.

The application of this principle for spectrum analysis is simple when the magnitude of the wave in question is a periodic function of time. The very fact that the wave is periodic is sufficient proof that the only frequencies that can be present in the wave are those corresponding to the basic repetition rate and its harmonics. Thus the frequency spectrum is confined to these specific frequencies and so it takes the form of a Fourier series. Knowing that the possible frequencies are restricted in this way, the problem of finding the frequency spectrum of a complex periodic wave is reduced to one of performing the above averaging process at each possible frequency. The period of the envelope of the Complex Wave is the proper time interval for averaging, and the integral formulation for obtaining this average is that for determining the coefficients in a Fourier series.

The principle holds equally well when the magnitude-time function is nonperiodic, but the concept is complicated by the fact that the frequency spectrum in such cases is transformed from one having a discrete number of components of harmonically related frequencies to one having a continuous band of frequencies.[6] Such spectra contain infinite numbers of sinusoidal

[4] The proper time interval is generally some integral multiple of the period corresponding to the difference in frequency of the two sinusoid waves.

[5] In practice it is generally necessary to multiply by both sine and cosine functions because of possible phase differences.

[6] One exception to this statement is the fact that any wave made up of two or more incommensurate frequencies is nonperiodic. Yet such waves will have a discrete spectrum if the number of components is finite. This incommensurate case is neglected throughout the discussion.

components, each of infinitesimal amplitude and so close together in frequency as to cover the entire frequency range uniformly.

The continuous band type of frequency spectrum is just as characteristic of non-periodic waves as the discrete spectrum is of periodic waves. This can be shown as a logical extension of the Fourier series representation of periodic waves. The transition from a frequency spectrum consisting of a series of discrete frequencies to one consisting of a continuous band of frequencies can be made by treating the non-periodic function as a periodic function in which the period is allowed to become very large. As the period approaches infinity the fundamental recurrence rate approaches zero, so that the harmonics merge into a continuous band of frequencies.

This does not of course change the basic realtionship between the frequency spectrum of a wave and its magnitude-time function. The magnitude-time function is still the sum of the components of the frequency spectrum. Also the frequency spectrum can still be obtained frequency by frequency, by averaging the product of the magnitude-time function and a unit sinusoid at each frequency. However, the actual transformations in the case of the non-periodic functions require summations over infinite bands of frequencies and over infinite periods of time and so fall into the realm of the Fourier and similar integral transforms.

However, in any case the problem of spectrum analysis reduces to an averaging process. The process can be performed by mathematical integration in all cases where a satisfactory analytical expression for the magnitude-time function is available. Fourier analysis provides a very powerful technique for setting up the necessary integrals in such cases.

This averaging process can also be done graphically. It is apparent from the theory that if the product of the magnitude-time function and the sinusoid is sampled at a sufficient number of points, spaced uniformly over the proper time interval, then the average of the samples gives the desired value. This technique is fully treated elsewhere[7] so that it will not be considered in detail here. However, use will be made of it in a qualitative way to augment the physical picture.

Non-Linear Aspects

The use of the frequency spectrum in transmission studies is generally limited to cases where the system in question is linear; that is, where the transmission is independent of the amplitude of the signal. However, the same techniques can still be used on systems employing successive linear and non-linear components, in cases where the transmission through the non-linear elements is independent of frequency. Under these conditions, the magnitude-time representation of the wave can be used in computing

[7] Whittaker and Robinson, Calculus of Observations.

the transmission over each non-linear section, where the transmission is dependent only on the amplitude, and the frequency spectrum used over each linear section, where the transmission is dependent only on the frequency. This a technique can be used on most pulse modulating systems because such non-linear elements as the modulators and limiters generally encountered are substantially independent of frequency.

Frequency Spectrum of the Single Pulse

The single pulse is a non-periodic function of time and so has a continuous frequency spectrum. In this case the Fourier transforms are simple. They are derived in Appendix A. Figure 1 gives a graphical representation of the magnitude-time function and the frequency spectrum of the pulse. The expressions are general and hold for pulses of any length or amplitude.

It is instructive to note that the frequency spectrum in this case can be

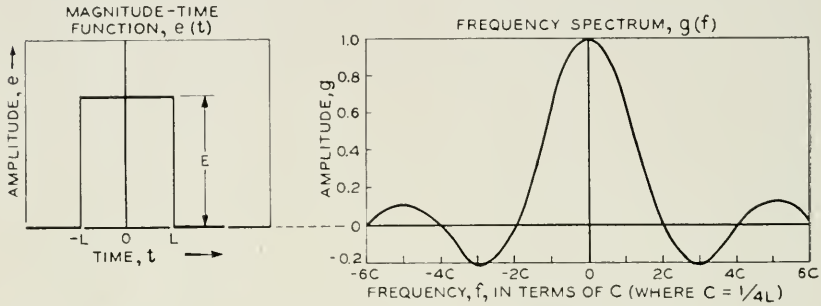


Fig. 1—Magnitude time and frequency spectrum representations of a single pulse.

determined by using the graphical technique mentioned previously. For example, consider the product of the magnitude-time function of the single pulse with a sinusoidal wave of given frequency and unit amplitude, so arranged in phase that its peak coincides with the center of the pulse. Theoretically the average of this product taken over the infinite period will give the relative magnitude of the component in the frequency spectrum of the pulse having the same frequency as the sinusoidal wave. In this case however, the average need only be taken over the length of the pulse, since the product vanishes everywhere else. Thus at very low frequencies, where the period of the sinusoidal wave is very much greater than the length of the pulse, the average is proportional to $2EL$ where E is the amplitude and $2L$ the length of the pulse. Then as the frequency increases, the average of the product, and hence the relative amplitude of the component in the spectrum, will first decrease. For the particular frequency such that the length of the pulse is one half the period, the relative amplitude will have

fallen to $2EL \times \frac{2}{\pi}$ ($\frac{2}{\pi}$ being the average value of a half wave of unit amplitude). Similarly when the frequency is such that the length of the pulse is a full wavelength, the average will vanish, and when the pulse length is one and a half times the wavelength, the average is negative, having two negative and one positive half waves over the length of the pulse, and the relative magnitude is $2EL \times \frac{2}{3\pi}$. These products are shown graphically on Fig. 2. Since these amplitudes correspond to those given in Fig. 1, for the spectrum components at $f = f_0 = 1/4L$, $2f_0$, and $3f_0$, it is apparent that the spectrum could be determined in this way.

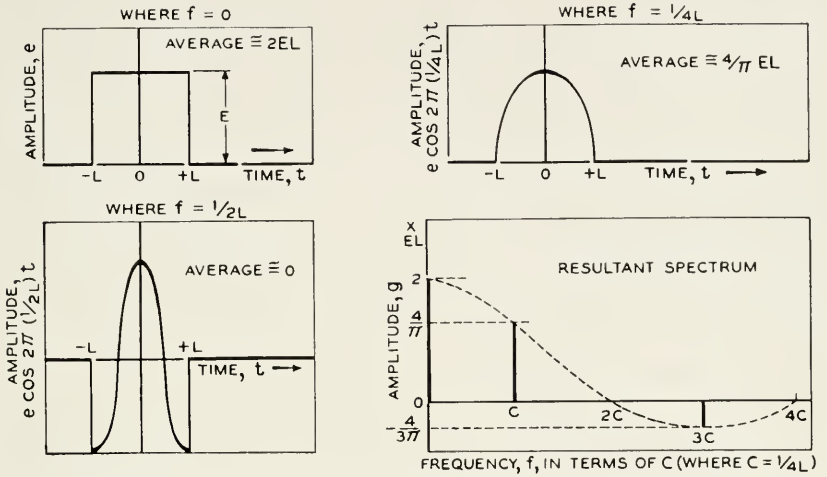


Fig. 2—Graphical derivation of spectrum of single pulse by averaging product of pulse with sinusoidal waves of various frequencies.

Basic Technique

In the analysis presented here, the single pulse and its spectrum will be used in such a way that the need for individual integral transforms for each complex wave form under study is avoided. The theory is simple.

A complex wave form may be approximated to any desired accuracy by a series of pulses, varying with respect to time in length, in amplitude, and in position. Now the spectra of these individual pulses are already known. Therefore, to find the frequency spectrum of the complex wave in question, it is necessary only to combine properly the spectra of the various pulses representing the complex wave.

Thus the process is theoretically complete. The procedure is first to

break down the given complex wave into a series of single pulses. Next the spectrum of each pulse is determined separately. Then the spectrum of the complex wave is obtained by combining the spectra of the various single pulses involved. One of the things to be demonstrated here is that it is perfectly feasible in many cases to perform these summations graphically, even though basically it does involve the handling of spectra each containing an infinite number of frequency components.

There are other wave forms that could be used as the fundamental building block instead of the single pulse. The unit step function is one possibility, since it is used in transient analysis for a similar purpose. However, the single pulse has obvious advantages when the complex wave to be analyzed is itself a series of pulses, as in pulse modulation. Again it would be nice to be able to choose as the fundamental unit a wave that has a discrete rather than a continuous band frequency spectrum, but it seems that any wave flexible enough to make a satisfactory building unit is inherently non-periodic and so has a continuous frequency spectrum. However the fact that the fundamental units have continuous spectra does not of itself complicate the results. If for example, the wave to be analyzed is periodic, the sum of the spectra of the various pulses must reduce to a discrete frequency spectrum. In the cases of interest here, when the pulse train under analysis is repetitive, combinations of identical pulses will be found to occur with the same fundamental period, and generally the first step in the summation of such spectra is to group the series of pulses into periodic waves with discrete spectra.

Manipulations of Single Pulses

In its use, the single pulse may be varied in amplitude, in length, and in position with respect to time. These changes have independent effects on the frequency spectrum. A variation in the amplitude of a pulse does not change its spectrum, except to increase proportionately the magnitudes of all components. A change in position of a pulse with time does not change the amplitude vs. frequency characteristic of the spectrum, but it does shift the phase of each component by an amount proportional to the product of the frequency and the time interval through which the pulse was shifted. A change in the length of a pulse will change the shape of the amplitude vs. frequency characteristic of the spectrum. Figure 3 shows this effect. However, if the center point of the pulse is not shifted in time, the relative phases of the components are not affected by such changes in length.

The single pulse can also be modulated to aid in the resolution of more complicated wave forms. This process is based on the use of the pulse as a function having a value of unity over a chosen time interval and a value of zero at all other times. Thus, to show a part of a sinusoidal wave, we need

only multiply this wave by a pulse of the correct length and proper phase with respect to the sinusoid to show only the desired piece of the wave. In this simple case it is not difficult to derive the spectrum because what are produced are the sum and the difference products of the modulating frequency with the spectrum of the pulse. This gives two single pulse spectra shifted up and down in frequency by the frequency of the modulation. An example of this is shown in Fig. 4, where the spectrum of a single half cycle is determined.

Pulse Position Modulation

For the first example, a simple form of pulse position modulation will be analyzed. The pulse train in this case is made up of pulses spaced T seconds

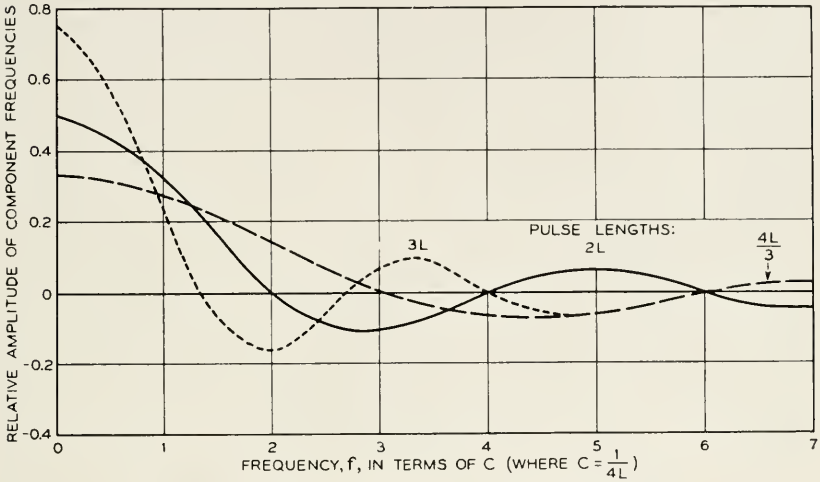


Fig. 3—Change in frequency spectrum with pulse length.

apart and the width of each pulse is a very small part of the spacing T. Such a pulse train is shown on Fig. 5. The pulse train is modulated by advancing or retarding the position (time of occurance) of the pulses by an amount proportional to the instantaneous amplitude of the signal at sampled instants T seconds apart. Figure 5 also shows the signal, in this case a sine wave of frequency $1/10T$, and the resulting modulated pulse train. The peak amplitude of the modulating sine wave is assumed to shift the position of a pulse by $1/4T$. The length and the amplitude of the pulses are the same since neither is affected in this type of modulation.

The first step in the analysis is to determine the spectrum of the pulse train before modulation. Each pulse contributes a spectrum of the form

shown on Fig 1. Now the phase of each component in such a spectrum is so arranged that the spectrum forms a series of cosine terms all of which have zero phase angle at the center of the pulse. From successive pulses T

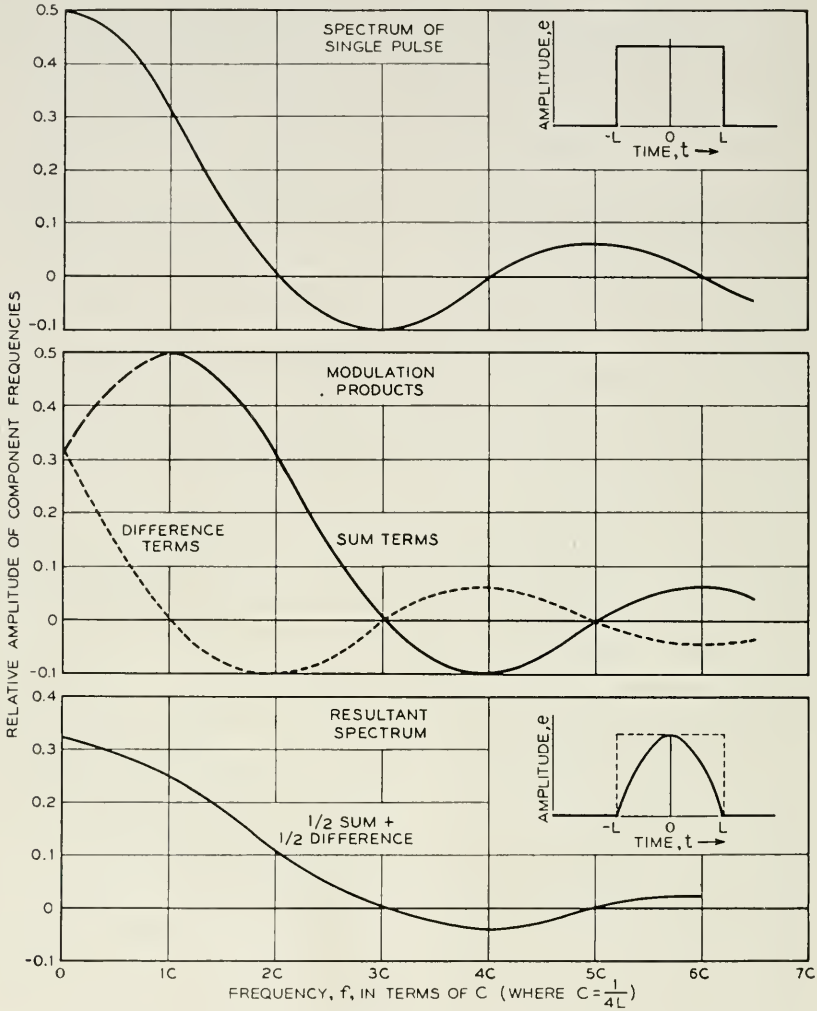


Fig. 4—Determination of spectrum of single half sine wave by modulation of single pulse spectrum with cos $2\pi ct$.

seconds apart, the component at any given frequency will have the same amplitudes, but the relative phases will be $2\pi fT$ radians apart. It is apparent that frequencies for which $2\pi fT$ is 2π or some multiple of 2π radians

apart, the contributions from all pulses add in phase. These are the frequencies nc, where $n = 1,2,3$ and $c = \frac{1}{T}$. It is also apparent that at frequencies for which the phase differences between the components are not an exact multiple of 2π radians apart, the contributions from enough pulses must be spread in phase over an effective range of 0 to 2π radians in such a way as to cancel one another. For example, take the particular frequency for which the difference in phase between pulses is 361° instead of 360°.

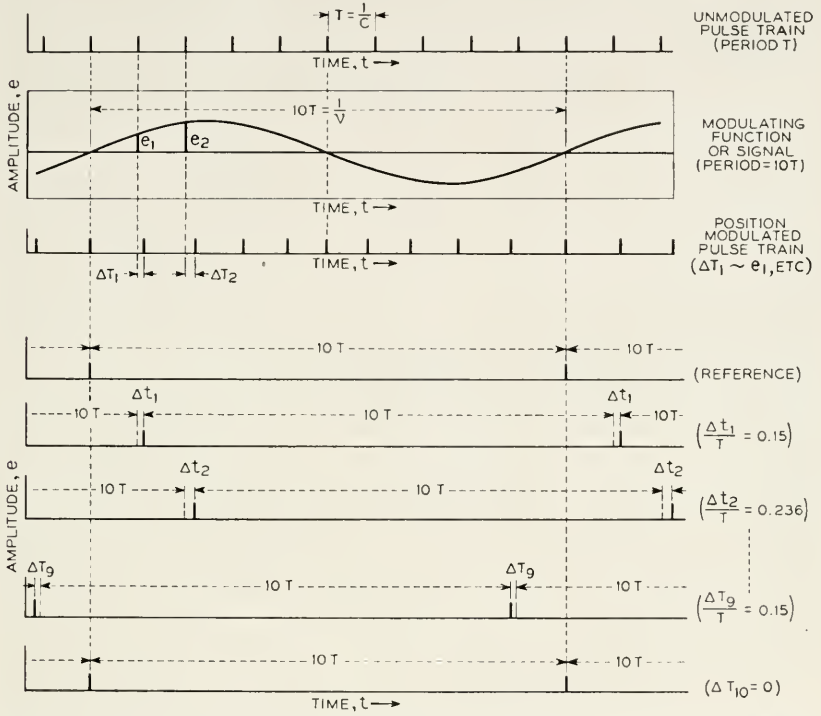


Fig. 5—Formation of pulse position modulated pulse train and its resolution into subsidiary unmodulated pulse trains.

The contribution from each preceding pulse will be effectively advanced in phase 1° with respect to its successor, so that the contributions from pulses 180 periods apart will be exactly 180° out of phase. Therefore over a sufficient number of pulses, the net contribution is zero.

The spectrum of the unmodulated pulse train is thus made up of a dc term plus harmonics of the frequency $C = 1/T$. The dc term is the average, and therefore is equal to $E \times 2L/T$, where E is the magnitude of the pulse. All of the other components have the same relative magnitudes that they have

in the single pulse spectrum. This gives a spectrum like that shown on Fig. 6. Figure 6 also shows for comparative purposes the spectrum of the subsidiary pulse wave consisting of every 6th pulse.

Thus in the unmodulated case, the pulses have a uniform recurrence rate and the resultant spectrum, found by adding those of the individual pulses, reduces to a train of discrete frequencies comprised only of the harmonics of the recurrence rate of the pulses. The fundamental frequency, correspond-

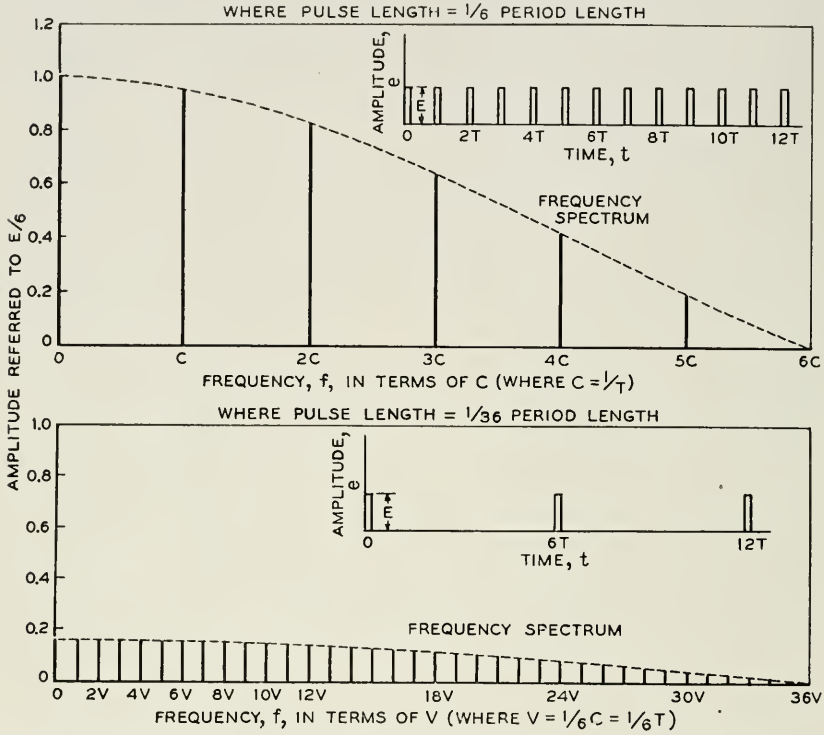


Fig. 6—Frequency spectrum of pulse trains where the spacing between the pulses is 6 and 36 times the pulse length respectively.

ing to the recurrence rate, and its harmonics will be called the carrier frequencies of the pulse train. The effect of modulating the pulse train is to modulate each of these carriers, producing sidebands of the signal about them.

When the pulse train is position modulated, the pulses are shifted in position by an amount ΔT, corresponding to the instantaneous amplitudes of the modulating function. The spectrum of each pulse is unchanged, since the pulse length remains constant. However, components of successive

pulses at the carrier frequency c and its harmonics will no longer add directly, because of the phase shifts that accompany the change in position. This phase shift is equal to ΔT, the shift in position, times the radian frequency of the component in question.

However, when the signal function is periodic, each pulse will have the same shift in position as any other pulse that occurs at the same relative instant in a later modulating cycle. Furthermore, when the carrier frequency is an exact multiple of the signal frequency i.e., $c = nv$, there will be a pulse recurring at the same relative instant in each cycle of v. Under these conditions, the pulse position modulated wave can be broken down into a group of unmodulated waves, each being made up of that series of pulses that recur at a given part of each modulating cycle, as shown in Fig. 5. These subsidiary waves are effectively unmodulated because, as each pulse recurs at the same instant in the modulating cycle, they are shifted to the same extent and hence will be uniformly spaced. This uniform spacing between pulses in a given wave is equal by definition to the period of the modulating function, and there will be as many of these unmodulated pulse trains as there are pulses in a single cycle. Thus, if $c = nv$, there will be n such pulse trains.

The reason for grouping the pulses into these unmodulated pulse tains is that unmodulated periodic trains have spectra of discrete frequencies. Since the pulse widths are all equal, and since the spacing between pulses is the same for each wave, the spectra of these unmodulated waves will all be identical. Furthermore, these spectra will be the same as that of the original carrier wave of pulses before modulation, except for two factors. First, the fundamental frequency is now v, corresponding to the modulating period, so that there are n times as many components as before. Secondly the amplitudes are reduced by the factor $\frac{1}{n}$ because there is only one pulse in these new waves to every n pulses in the original wave. Thus, instead of having a spectrum made up of the carrier frequency and its harmonics, we now have one made up of harmonics of v. Since $c = nv$, such frequencies as c, c, $\pm v$, $c \pm 2v$, etc., are included. An example of the spectra of both the subsidiary and original pulse waves is shown on Fig. 6, for the case where $n = 6$.

Thus the problem of finding the spectrum of such a pulse position modulated wave is reduced by this procedure to adding up the n equal components at each of the frequencies of interest, such as c and $c \pm v$, allowing for the phase difference between components corresponding to the position of one pulse with respect to that of the other n–l pulses in one modulating cycle. As an example, suppose $n = 10$ and the frequency to be computed is $c + v$. Now $c + v$ is 10% higher in frequency than c. Thus in the unmodulated

case, when the n pulses are equally spaced, they are 360° apart at c and consequently 360° + 36 or 396° at $c + v$. Therefore in the unmodulated case, each component would be advanced in phase 36° with respect to the previous one, so that the diagram of the 10 components would form the

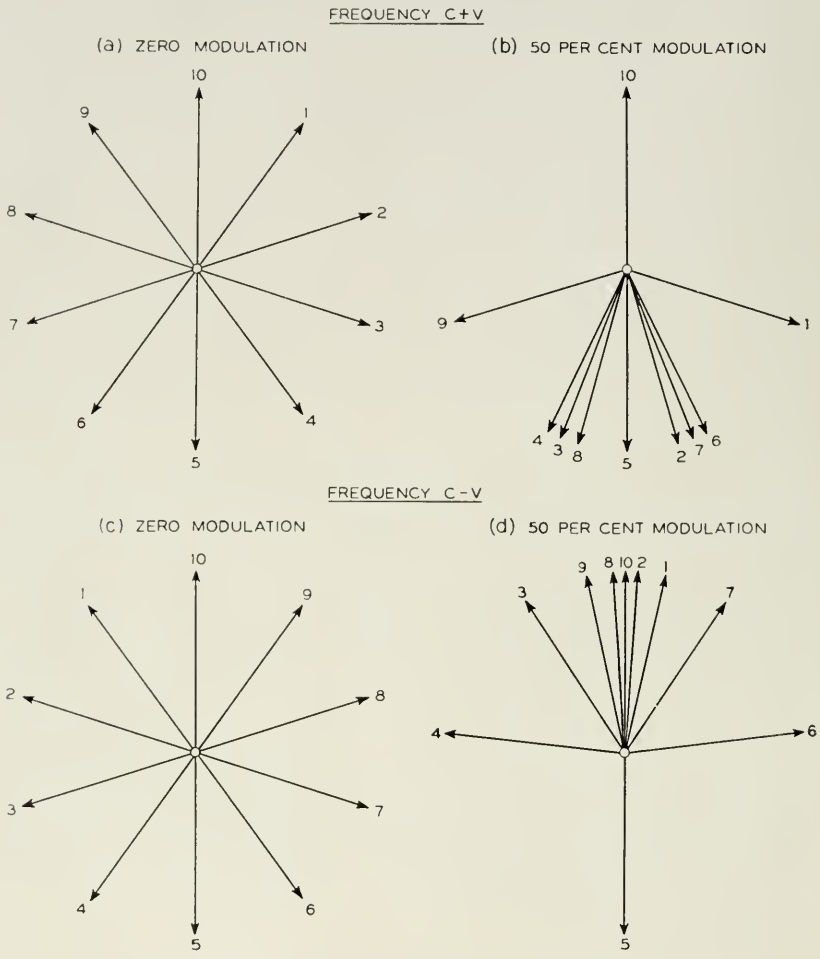


Fig. 7—Vector pattern of subsidiary pulse components.

vector pattern shown on Fig. 7A. The successive components are numbered 1 to 10. The sum in this unmodulated case is of course zero.

Now the effect of modulation is to shift the relative phases of these components by an amount determined by the shift in position of the corresponding pulses. When these relative phase shifts are such as to spoil the can-

cellation of the 10 components, a net component of this frequency is produced in the frequency spectrum of the pulse wave. Taking the example shown in Fig. 5, the 10 components in Fig. 7A would be shifted to the positions shown in Fig. 7B. These shifts in relative phase are determined in the following way. Figure 5 shows that the number 1 pulse is retarded an amount ΔT_1 equal to 15% of T, the normal spacing between pulses. Thus at the carrier frequency c, the phase shift between the component from this retarded pulse and the reference pulse is 15% more than 360° or 414°. Thus the component at the carrier frequency c from the first subsidiary pulse train is shifted 54° from its unmodulated position.

At $c + v$, since the frequency is 10% higher, the net shift is 10% more than at c or 59.5°. Thus the number 1 component on the vector diagram of Fig. 7B is rotated 59.5° clockwise from its unmodulated position shown on Fig. 7A.

Similarly pulses 2 and 3 are each shifted in position by equal amounts, ΔT_2 and ΔT_3. These shifts in position give 85° phase shift at the carrier frequency. Hence components 2 and 3 at $c + v$ are each rotated 10% more or 93.5° from their respective unmodulated reference positions shown on Fig. 12A. Component number 4 is shifted 59.5° clockwise just as number 1. Component 6 and 9 are also shifted 59.5° each, but in this case the modulating function has the reverse polarity so that the components are rotated counterclockwise. Similarly components 7 and 8 are rotated 93.5° counterclockwise.

The sum of these components in the vector diagram of Fig. 7B gives a resultant that is negative with respect to the reference direction and the magnitude that is 58% of the reference magnitude, where the reference magnitude and direction are those for the carrier c with no modulation.

This gives the relative magnitude and phase of the $c+v$ term produced by pulse position modulation for the case where the modulating function is a sine wave of frequency $v = c/10$ with a peak amplitude just large enough to shift a pulse by 1/4 of T, where T is the spacing between unmodulated pulses. A shift of this magnitude will be defined here as 50% modulation on the basis that 100% modulation should be $1/2T$, the maximum displacement that can be used without possible interference between pulses.

In the same way the other component frequencies in the spectrum such as $c, c - v, c \pm 2v$, etc., have been computed for the above case of 50% modulation, and for other peak amplitudes of the modulating sine wave giving 25%, 70% and 100% modulation. In all cases the frequency of the modulating function was held at $v = c/10$. This information is plotted on Fig. 8, showing v, c and the various components of the frequency spectrum that represent the sidebands about the carrier frequency c, as a function of the peak % modulation.

The above solution assumed a special case where c was an exact multiple of v. The purpose of this assumption was to simplify the problem to the extent that the periodicity of the modulated wave would be the same as that of the modulating function. There are two other possible cases. For one, the ratio of c to v could be such that a pulse would occur at the same instant of the modulating period only once every so many periods. The actual periodicity of the modulated pulse wave would be reduced accordingly because it would make the same number of periods of the modulating function before the modulated pulse train is repeated. This is a result of the fact that pulse modulation provides for a discrete sampling rather than a continuous measure of the modulating wave. The technique of spectrum analysis demonstrated above is just as applicable to this case as it was to the simpler one. However, there will be comparatively more terms to be handled. The other possible case is the one where c and v are incommensurate.[8] In this case, the resulting modulated wave is non-periodic. However, on the basis that the spectrum is practically always a continuous function of the signal frequency, this case has received no special attention here.

At frequencies for which c is very much greater than v, so that the number of component pulse trains becomes too numerous to handle conveniently in the above fashion, the sidebands about each carrier or harmonic of the switching frequency can be computed by the standard methods for phase modulation, as the next section will demonstrate. This result follows directly from the theorem that as the carrier frequency c becomes large with respect to v, pulse position modulation merges into a linear phase modulation of each of the carriers.

Pulse Position Modulation vs Phase Modulation

When a pulse, in a pulse position modulated wave, is shifted by 1/2 the spacing between pulses (100% modulation) it is apparent from the previous discussion that the component of the carrier in the frequency spectrum of the pulse is shifted by 180°. Therefore to compare the spectrum of a pulse position modulated wave like that on Fig. 8 with the equivalent spectrum of a phase modulated wave, what is needed is Fig. 9, showing the frequency spectrum of a phase modulated wave of the form $Cos(ct - k \sin vt)$ as a function of k for values of k up to π radians or 180°. The computation of the frequency spectrum of such a phase modulated wave has been adequately covered elsewhere and all that is done here is to give the brief development shown in appendix B.

[8] Mr. W. R. Bennett has pointed out that this incommensurate case is the general one. It requires a double Fourier series, which reduces to a single series when the signal and carrier frequencies are commensurate. This analysis is based on the single Fourier series.

A comparison of the spectra on Figs. 8 and 9 shows that the sidebands have the same general pattern. However comparative sidebands are not

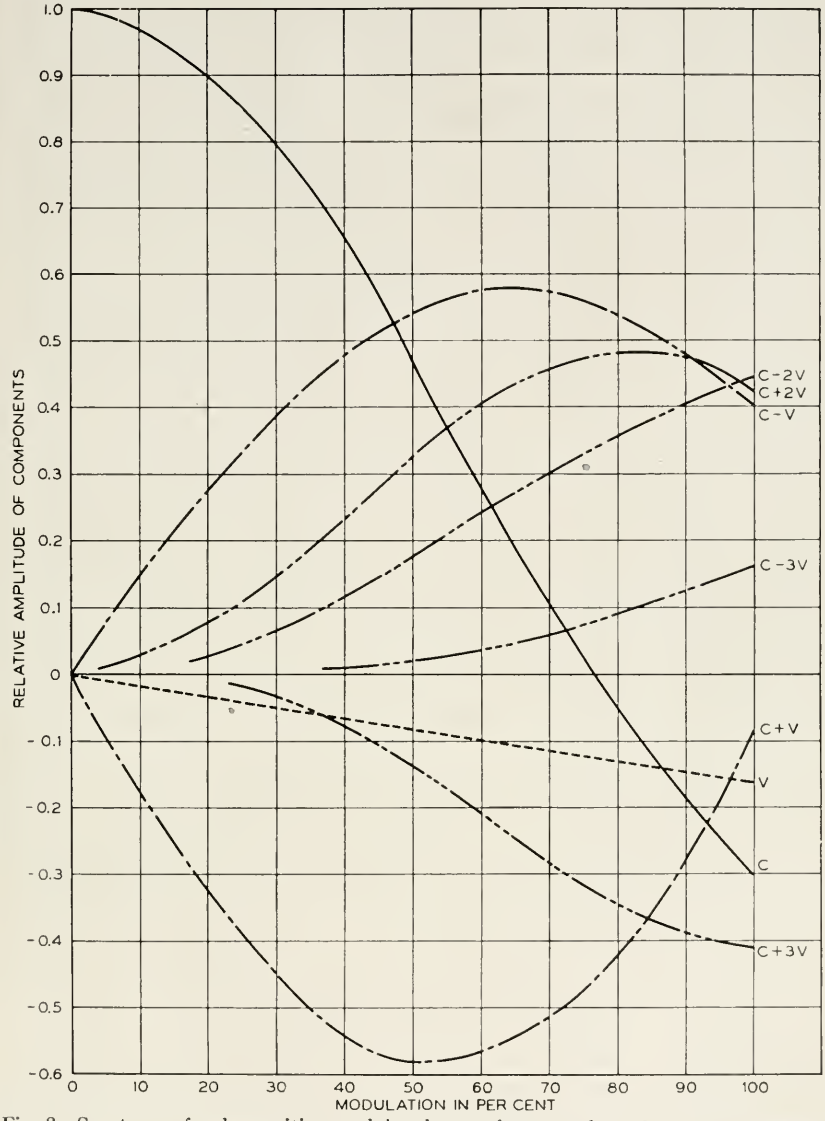


Fig. 8—Spectrum of pulse position modulated wave for case where the carrier frequency C is 10 times the signal frequency v.

quite equal in the two cases. In fact comparable upper and lower sidebands in the case of the pulse modulated wave shown on Fig. 8 are not

equal in absolute magnitude to each other. This lack of symmetry is due to the fact that c is only 10 times v.

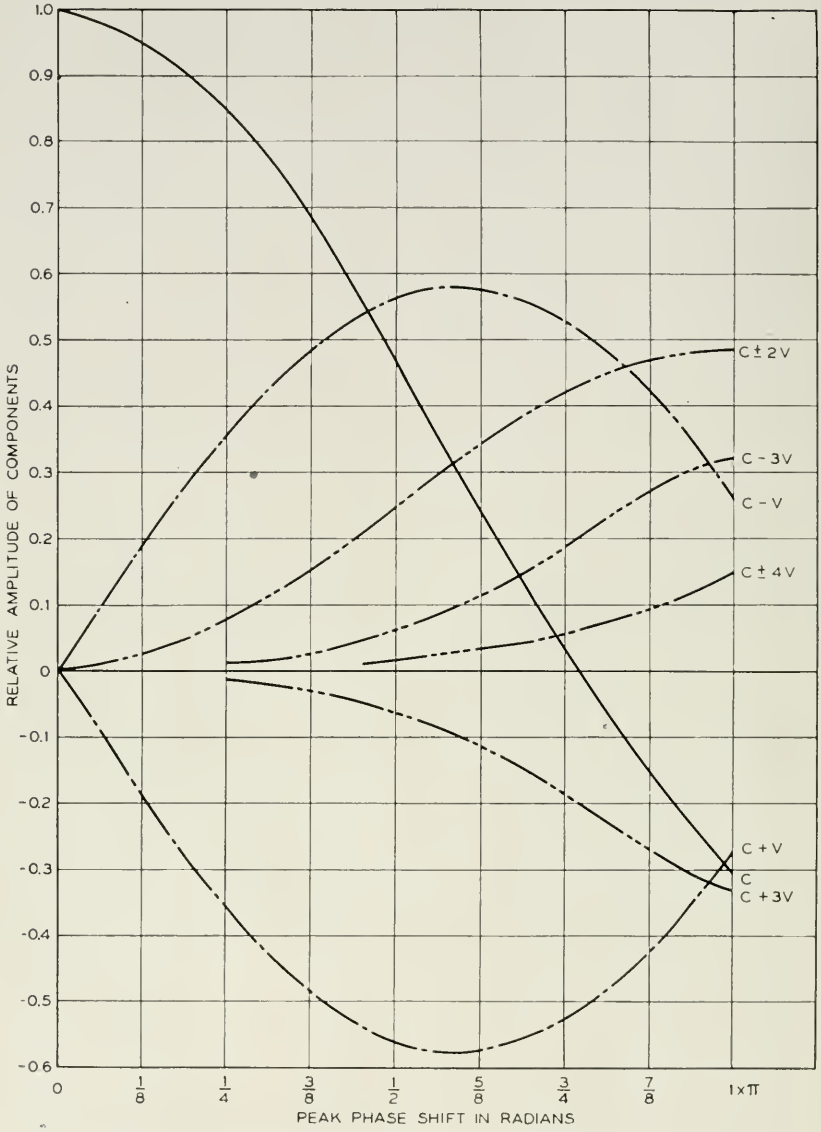


Fig. 9—Spectrum of phase modulated wave cos $(ct + k \sin vt)$ as function of peak phase shift k for values of k up to π radians.

One way of proving this is to go through the process of computing the $c - v$ term in this pulse modulated wave just as the $c + v$ term was computed

earlier. Since the frequency $c - v$ is 10% less than c, the unmodulated pattern of the 10 subsidiary components, as shown on Fig. 7C, is the mirror image of that for $c + v$ in 7A, for the first component is now 360° less 10% or 324°, and subsequent components are each retarded 36° with respect to the previous one. When the pulse train is modulated the effect is similar to the case for $c + v$ and, for the same per cent modulation, the Vector pattern of Fig. 7D is formed. The resultant in this case differs from that of 7B in sign as well as in magnitude. The difference in sign comes from the fact that, since component 1 in 7A corresponds to component 9 in 7C and component 2 in 7A to component 8 etc., the modulation in the case of $c - v$ rotates these corresponding components in opposite directions. The difference in magnitude is due to the fact that since $c - v$ is an appreciabley lower frequency than $c + v$ in this case (approx. 20%), the phase shift corresponding to a given shift in pulse position is proportionately less. Thus the corresponding Vector components are not shifted the same number of degrees. Thus the absolute magnitudes of $c + v$ and $c - v$ are not equal in this case.

It is apparent that this difference in magnitudes of $c + v$ and $c - v$ becomes smaller as the carrier frequency c becomes larger with respect to v. In the limiting case of c very much greater than v, $c + v$ and $c - v$ would each be shifted the same number of degrees as c itself. If this more or less compromise shift of c is used to compute the $c \pm v$, $c \pm 2v$, and c $\pm$ $3v$ terms, then the resulting frequency spectrum is that of the phase modulated carrier on Fig. 9.

The higher harmonics of c in the pulse position wave are similarly phase modulated and the interesting point is that $2c$ is modulated through twice as many degrees phase shift and $3c$ 3 times as many degrees, etc. Thus a single pulse position modulator could be designed to produce a harmonic of c with almost any desired degree of phase modulation. This is a useful method for obtaining a phase modulated wave, or with a 6 db per octave predistortion of the signal, a frequency modulated wave.

Figure 8 also shows a term in v itself, which has been neglected so far in the discussion. It is apparent that the components at v contributed by the 10 subsidiary unmodulated waves must form the same kind of vector pattern as those of $c + v$ in Fig. 7. However, in this case $c + v$ is eleven times v in frequency, so that the components of v are rotated only one eleventh as much for a given pulse diplacement. Thus the magnitude of v at 100% modulation is equal to that of $c + v$ at approximately 9% modulation. For different frequency ratios of c to v the relationship of the v term to $c + v$ will vary, and it is apparent that for c very much greater than v, the v term will vanish. The relationship is such that the amplitude of the v component out of the modulator at a given per cent modulation is directly proportional to its own frequency v for all frequencies less than approximately one quarter

of c, and the phase is 90° with respect to the input. Thus the modulator puts out a signal component that is the derivative of the input signal.

To summarize the case of pulse position modulation, the frequency spectrum may be determined by the methods based on subdividing the modulated pulse train into a series of unmodulated ones when the ratio of c to v is small, and by treating each harmonic of the carrier as a phase modulated wave of the form $Cos\ n\ (ct + \theta)$, where θ is the modulating function, when the ratio of c to v is large. In the case treated here, the modulating function was a simple sinusoidal wave. Of course the analysis holds for more complicated wave shapes having frequency spectra of their own. In this event however the restriction on the relative magnitudes of the frequencies v and c should be taken as one on c and the highest frequency in the modulating spectrum. The complexity of the modulating function does not affect the analysis when it is done by this technique of subdividing the pulse train, since all that need be known is how much each pulse is shifted, and this can be done graphically. The analysis given here has neglected the length of the individual pulses. This was done when it was assumed that the individual contributions from the various pulse trains had the same amplitude at all frequencies. For any finite pulse width, the relative magnitudes of the various components must be modified by the $\frac{sin\ x}{x}$ factor of the single pulse, as shown on Fig. 6.

As mentioned in the introduction, a complex wave could be analyzed by multiplying its magnitude-time characteristic by unit sinusoids at each frequency in question, sampling the product at a sufficient number of points uniformly spaced over a cycle of the envelope of the complex wave, and then averaging the values of the product thus obtained. This technique is particularly applicable to the analysis of pulse position modulated waves since, by taking the centers of the pulses of the modulated wave as the sampling instants, it is possible, with a finite number of samples (same as the number of pulses) to get the same results as though a very much greater number of uniformly spaced samples were taken. The interesting thing to note here is that the actual computations that would be involved in applying this sampling method of analysis to a pulse position modulated wave are almost identically the same calculations as required by the technique of resolving the pulse train into unmodulated subsidiary pulse trains used here.

Pulse Width Modulation

Pulse Width Modulation as defined here could also be termed "pure" pulse length modulation. The pulse train in the reference or unmodulated condition is a recurrent square wave, and the lengths of the pulses will be varied by the modulation without changing the position of the centers of the pulses. The term "pure" pulse length modulation is applicable to this

special case where the phase relationship between spectra of adjacent pulses does not change with modulation because the centers of the pulses are not shifted by the modulation. The conventional form of pulse length modulation, where one end of the pulse is fixed in position, combines both this pulse width modulation and the pulse position modulation previously analyzed. The interest in this case of pulse width modulation arose in connection with the analysis of "hunting" servomechanisms, and the analysis provides a basis for a general solution of the response of a two-position switch or ideal limiter to various forms of applied voltages.

Since the unmodulated wave is a square wave with pulses of length $2L$ recurring at intervals of $T = 4L$, it has the familiar square wave spectrum including a d-c term, a fundamental term or carrier of frequency $c = 1/T$, a 3rd harmonic with a negative amplitude $1/3$ that of the fundamental, etc. Figure 10 shows clearly that this spectrum is the sum of single pulses of width $2L$ spaced $T = 4L$ seconds apart. In the summation, all frequencies cancel except harmonics of c and, since they all add directly in phase, the component frequencies in the resultant spectrum have the same relative amplitudes as they have in one single pulse.

When this pulse train is modulated, the width of each pulse becomes $2(L + \Delta L)$, where the magnitude of ΔL depends in some specified way on the magnitude of thhe modulating function at the instant corresponding to the center of the pulse. For simplicity, the case will be taken where ΔL is proportional to the magnitude of the modulating function. For 100% modulation, ΔL will be assumed to vary from $-L$ to $+L$. Figure 3 shows how the relative amplitude of the components of the frequency spectrum of a pulse vary for 3 different values of ΔL , along with the equation that governs these amplitudes.

If the modulating function has a periodicity v such that $c = 10v$, then every 10th pulse, recurring at the same instant in each modulating cycle, will be widened to the same extent and so can be formed into a subsidiary unmodulated pulse train, as was done on Fig. 5 for the pulse position modulated wave.

Again vector diagrams like those in Fig. 7 may be formed showing the contribution of each of these subsidiary pulse trains at various frequencies such as c, $c + v$ and $c - v$. When the waves are unmodulated, the vector diagrams for the same frequencies will be the same as those for the pulse position modulated case, except for the absolute amplitudes of the components, as long as $c = 10v$ in each case. When the pulse width system is modulated, however, the modulation does not rotate the individual vector components as in the pulse position case since the spacing between pulses is not changed. What the pulse width modulation does is to change the length of the individual component vectors exactly as it does in the case of

the single pulses shown on Fig. 3. This change of magnitude, of course, can spoil the cancellation of the ten unmodulated components at some frequency like $c + 2v$ just as effectively as rotating them did in the case of the pulse position modulated wave, thus producing a spectrum component at that frequency.

As an example, the case will be taken where the modulating function is a

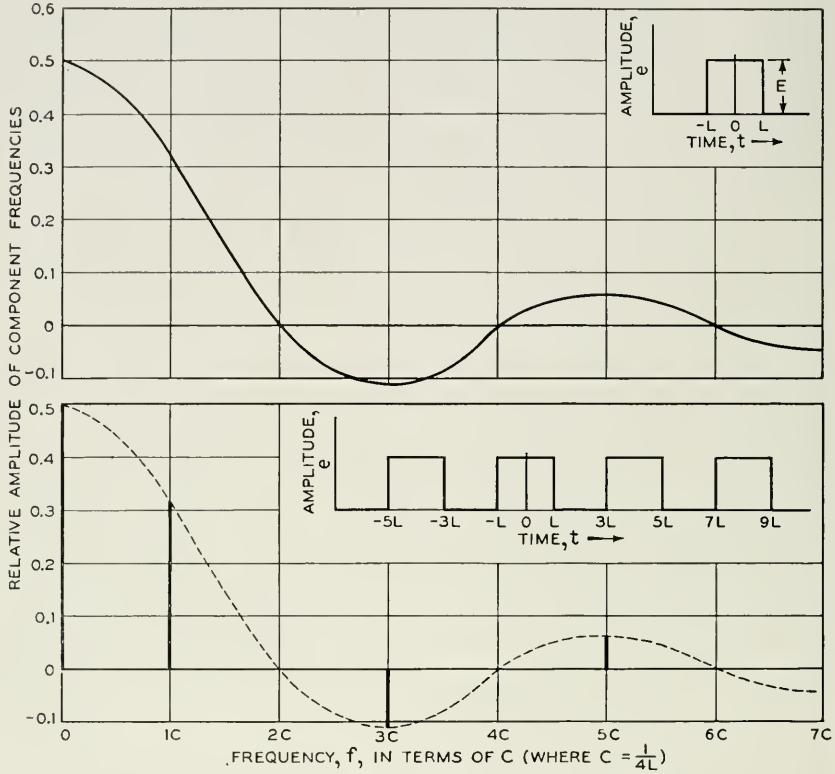


Fig. 10—Comparative spectra of square wave and single pulse.

sinusoid of frequency v. Then the change in width with modulation is given by the formula

$$\frac{\Delta L}{L} = k \sin vt.$$

Since $c = 10v$, the successive subsidiary pulse trains will be modulated an amount $\left(\frac{\Delta L}{L}\right)_m = k \sin\left(2\pi \frac{m}{10}\right)$ as m takes on the values from 1 to 10. Thus the spectra of these subsidiary pulse trains with pulses of length $2(L +$

ΔL_m) recurring every $1/v$ seconds will be a Fourier series of harmonics of v. The amplitude of the nth term of this series will be

$$B_n = \frac{2E}{10\pi n} \sin\left[\frac{\pi n}{2}\left[1 + k \sin\left(\frac{2\pi n}{10}\right)\right]\right].$$

This expression may be found from appendix C, equation (5a). Combining

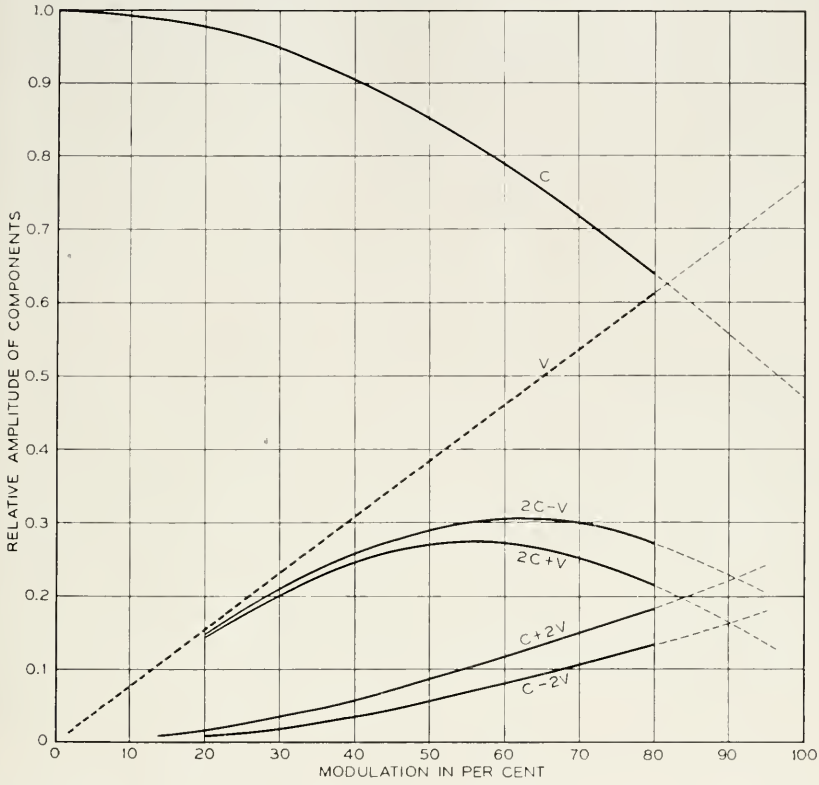


Fig. 11—Spectrum of pulse width modulated wave for case where carrier frequency C is 10 times the signal frequency v.

the 10 such components at each frequency, as shown on Fig. 7 for the case of the pulse position modulated wave, the spectrum for this case of Pulse Width Modulation on Fig. 11 is produced. This spectrum is comparable to that on Fig. 8 for the pulse position modulated case.

Pulse Width vs Amplitude Modulation

That pulse width modulation is a form of amplitude modulation of the carriers of the unmodulated pulse train is shown mathematically by Equa-

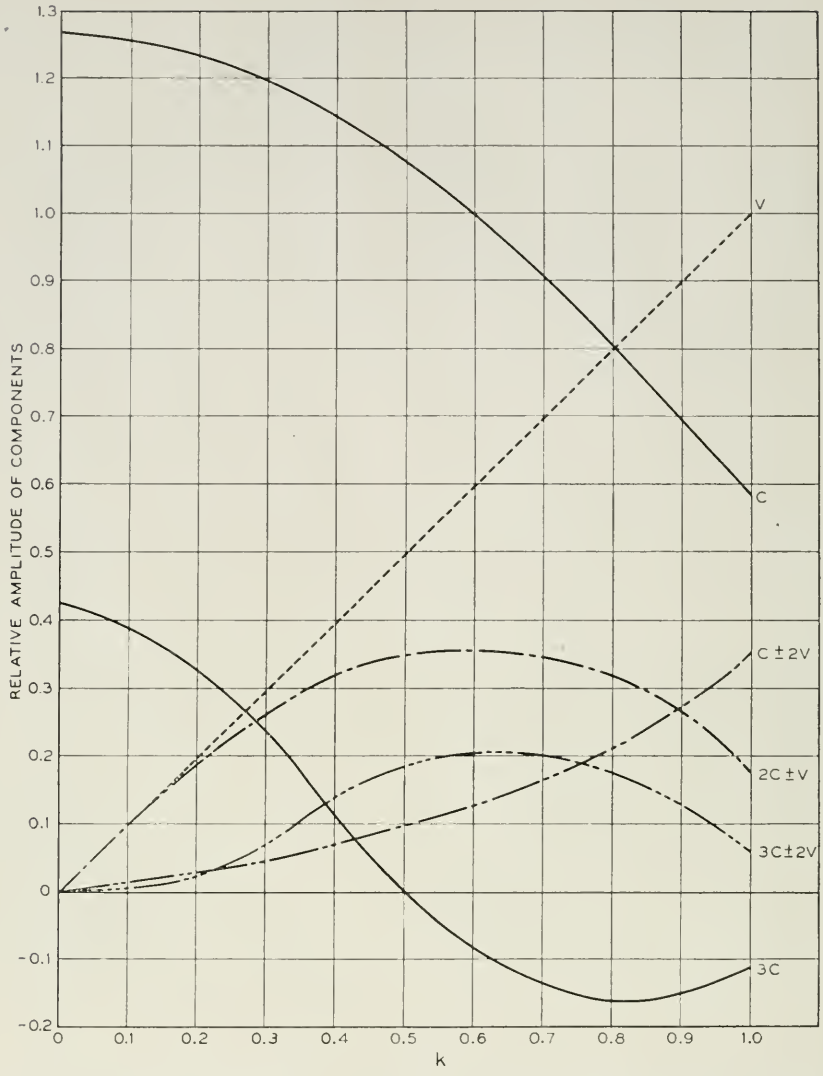


Fig. 12—Response of ideal limiter to simultaneously applied isosceles triangle wave and sine wave inputs. k is the ratio of the peak amplitudes of sinusoidal and triangular waves at the input.

tion (8) in Appendix C, where the spectrum is developed as a Fourier series in harmonics of the pulse rate c with the modulation affecting only the amplitude of the coefficients.

This mathematical analysis is continued in Appendix D where the fre-

quency spectrum is determined for $\frac{\Delta L}{L} = k \sin vt$. The spectrum thus computed is shown in Fig. 12.

An example of this type of pulse modulator is given by a two position switch or ideal limiter when the signal to be modulated is applied simultaneously to the limiter with an isosceles triangle wave as carrier. The carrier should have a higher peak amplitude than the signal and a recurrence rate based on the desired carrier frequency. Figure 12 is arranged to show the output spectrum for such a limiter in terms of k, when k is the ratio of the peak amplitudes of the sinusoidal signal and triangular carrier wave inputs.

A comparison of this spectrum with that on Fig. 11 shows that the two spectra have almost the same form. c and v have the same amplitude characteristics in each case. The $c \pm 2v$ and $2c \pm v$ terms have differences that are like those found before in comparing the pulse position modulated wave on Fig. 8 and the phase modulated carrier on Fig. 9. As in that case, when c becomes very much greater than v the differences vanish.

Application of Pulse Width Modulator

Practical interest in this case lies in the fact that the signal is present in the output spectrum with a linear characteristic that makes such a modulator a linear amplifier. The "on-off" or "hunting" servomechanism is based on a modified form of such an amplifier in which the carrier is supplied by the self oscillation of the system. The term modified form is used because the self oscillations in general are more nearly sinusoidal than triangular in form and so do not give a linear change in pulse length over as wide a range of input amplitudes as does a triangular carrier. No attempt will be made to analyze such a system here since it has been handled elsewhere.[9] However the above method is applicable to such problems regardless of the shape of the carrier or the signal.

Other Forms of Pulse Modulation

Another form of pulse modulation of interest is that of pulse length modulation in which either the start or the end of each pulse is fixed, so that the centers of the pulses vary in position with the length. This is a combination of both the pulse position and the pulse width modulations described above and can be analyzed by a combination of the methods developed.

These same methods are also applicable to the analysis of frequency and phase modulated waves after they have been put through a limiter, as they generally are before detection.

[9] See L. A. Macall, "The Fundamental Theory of Servomechanisms" D. Van Nostrand Company, 1945.

APPENDIX A

Fourier Transforms For Single Pulse

The amplitude $g(f)$ of the component of frequency f in the spectrum of the Complex Magnitude-time function $e(t)$ is given by the d-c component of the Modulation products of $e(t)$ and $\cos 2\pi ft$, found by averaging the product over the period of the complex wave.

Thus, for non-periodic waves, where the period is from $-\infty$ to $+\infty$, the amplitude of the spectrum at f is

$$g(f) \cong \int_{-\infty}^{\infty} e(t) \cos 2\pi ft \, dt. \tag{1}$$

For the single pulse, where $e(t) = E$ for $-L < t < L$ and $e(t) = 0$ for all other values of t, equation (1) reduces to

$$g(f) \cong \int_{-L}^{L} E \cos 2\pi ft \, dt. \tag{2}$$

Integrating,

$$g(f) \cong \frac{E}{2\pi f} \sin 2\pi ft \Big|_{-L}^{L}$$

or

$$g(f) \cong \frac{E}{\pi f} \sin 2\pi fL. \tag{3}$$

Equation (3) is the expression for $g(f)$ plotted on Fig. 1.

Similarly, in the case of the single pulse, each increment in frequency df contributes a factor proportional to $g(f) \cos 2\pi ft \, df$ to the composition of $e(t)$, so that

$$e(t) = \int_{-\infty}^{\infty} g(f) \cos 2\pi ft \, df. \tag{4}$$

Substituting in (4) the expression for $g(f)$ given by equation (3), this becomes

$$e(t) \cong \frac{E}{\pi} \int_{-\infty}^{\infty} \frac{\sin 2\pi fL}{f} \cos 2\pi ft \, df. \tag{5}$$

APPENDIX B

Frequency Spectrum Of Phase Modulated Wave

The Phase Modulated Wave in this case is given by

$$\cos (ct - k \sin vt) = \cos (ct) \cos (k \sin vt) + \sin (ct) \sin (k \sin vt)$$

Now $\cos (ct) \cos (k \sin ct) = J_0 (k) \cos (ct)$

$$+ J_2 (k) \cos (c - 2v) t$$

$$
\begin{aligned}
& & &+ J_2(k) \cos (c + 2v)\, t + \cdots \\
\text{and } \sin (ct) \sin (k \sin ct) & &&= J_1(k) \cos (c - v)\, t \\
& & &- J_1(k) \cos (c + v)\, t \\
& & &+ J_3(k) \cos (c - 3v)\, t \\
& & &- J_3(k) \cos (c + 3v)\, t + \cdots \\
\therefore \cos (ct - k \sin vt) & &&= J_0(k) \cos (ct) \\
& & &+ J_1(k) \cos (c - v)\, t \\
& & &- J_1(k) \cos (c + v)\, t \\
& & &+ J_2(k) \cos (c - 2v)\, t \\
& & &+ J_2(k) \cos (c + 2v)\, t \\
& & &+ J_3(k) \cos (c - 3v)\, t \\
& & &- J_3(k) \cos (c + 3v)\, t + \cdots
\end{aligned}
$$

APPENDIX C

In this Appendix the spectrum of a train of rectangular pulses of length $2(L + \Delta L)$ recurring every T seconds, will be found from the spectrum of a single pulse of this train.

For the single pulse at any frequency f,

$$g(f) \cong \frac{E}{\pi f} \sin 2\pi f(L + \Delta L). \tag{1}$$

For a series of such pulses recurring with a spacing $T = 1/c$, then the sum of spectra of the individual pulses form a Fourier series of harmonics of c. Thus

$$e(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos 2\pi nct, \tag{2}$$

where A_n is the sum of an infinite number (one from each pulse) of infinitesimal terms $g(nc)$ and $g(-nc)$, shown in (1). Thus

$$A_n \cong 2\Sigma \frac{E}{\pi nc} \sin 2\pi nc(L + \Delta L) \tag{3}$$

Now to put an absolute value to the amplitudes $g(f)$ shown in equation (1), it is necessary to average them over the recurrence period of the single pulse, making them infinitesimals. However, in the train of pulses recurring every $T = 1/c$ seconds, the amplitude of A_n can be determined by averaging the terms in (1) over an interval T. Then

$$A_n = \frac{2E}{\pi ncT} \sin 2\pi nc(L + \Delta L). \tag{4}$$

When $T = 4L = 1/c$, (4) reduce to

$$A_n = \frac{2E}{\pi n} \sin \frac{n\pi}{2} \left(1 + \frac{\Delta L}{L}\right) \tag{5}$$

For the example taken in the text, when the pulse train was subdivided into 10 subsiding pulse trains, the period $T = 1/v = 10/c = 40L$. Thus in this case, the Fourier coefficients of the harmonics of v are

$$B_n = \frac{2E}{10\pi n} \sin \frac{\pi n}{2} \left(1 + \frac{\Delta L}{L}\right). \tag{5a}$$

The expression for A_n in equation (5) can be put in simpler form by using the formula for the sin of the sum of two angles. In this way, we get

$$A_n = \frac{2E}{\pi n} \left[\sin \left(\frac{\pi n}{2}\right) \cos \left(\frac{\pi n}{2} \frac{\Delta L}{L}\right) + \cos \left(\frac{\pi n}{2}\right) \sin \left(\frac{\pi n}{2} \frac{\Delta L}{L}\right) \right]. \tag{6}$$

Now, for n odd, $\sin \frac{\pi n}{2}$ alternately assumes the value ± 1 and $\cos \frac{\pi n}{2}$ vanishes, and for n even, $\cos \left(\frac{\pi n}{2}\right)$ alternately assumes the value ± 1 and $\sin \frac{\pi n}{2}$ vanishes. The A_0 term, being the d-c average of the pulse train, is given by

$$\frac{E/2(L + \Delta L)}{T} = \frac{E}{2} \left(1 + \frac{\Delta L}{L}\right). \tag{7}$$

If the pulse train is transformed by shifting the zero so that it alternates between $\pm E/2$ instead of O and E, the first term in equation (7) vanishes and (2) becomes, from (6) & (7),

$$\left.\begin{aligned} e(t) &= A_0 + A_1 \cos 2\pi ct \\ &\quad + A_2 \cos 2\pi\, 2ct + \cdots \end{aligned}\right\}$$

Where

$$\left.\begin{aligned} A_0 &= \frac{E}{2} \left(\frac{\Delta L}{L}\right) \\ A_1 &= \frac{2E}{\pi} \cos \left(\frac{\pi}{2} \frac{\Delta L}{L}\right) \\ A_2 &= \frac{2E}{2\pi} \sin \pi \left(\frac{\Delta L}{L}\right) \\ A_3 &= \frac{2E}{3\pi} \cos \frac{3\pi}{2} \left(\frac{\Delta L}{L}\right) \end{aligned}\right\} \tag{8}$$

etc.

APPENDIX D

The purpose of this section is to compute the spectrum of the carrier given by equation (8) in Appendix C as their amplitudes vary with $\frac{\Delta L}{L} = k \sin vt$.

For the d-c term,

$$A_0 = \frac{E}{2}\frac{\Delta L}{L} = \frac{E}{2} k \sin vt.$$

For the fundamental or c term,

$$A_1 \cos 2\pi ct = \frac{2E}{\pi} \cos\left(\frac{\pi}{2} k \sin vt\right) \cos 2\pi ct$$

Using the Bessel's expansion of cos (2 sin θ), we get,

$$A_1 \cos 2\pi ct = \frac{E}{2\pi} \begin{cases} J_0(k) \cos 2\pi c \\ +J_2(k) \cos 2\pi(c - 2v)t \\ +J_2(k) \cos 2\pi(c + 2v)t \\ +\cdots \text{ etc.} \end{cases}$$

In a similar fashion, the other terms can also be computed, giving the spectrum shown on Fig. 12, where $J_0(k)$ becomes the amplitude of c, $J_2(k)$ the amplitude of either $c + 2v$ or $c - 2v$, etc.

Abstracts of Technical Articles by Bell System Authors

Commercial Broadcasting Pioneer. The WEAF Experiment: 1922–1926.[1] William Peck Banning. *WEAF*, the radio call letters which for nearly a quarter of a century designated a broadcasting station famous for its pioneering achievements, ceased last November to have its old significance. *WNBC* are the new call letters. This book is an excellent record of the four years during which this station was the experimental radio broadcasting medium of the American Telephone and Telegraph Company.

The author indicates that the WEAF experiment aided the development of radio broadcasting in three ways:

First, in the scientific and technological field.

Second, in the emphasis of a high standard for radio programs.

Third, in determining the means whereby radio broadcasting could support itself.

When *WEAF* changed hands from the American Telephone and Telegraph Company to new ownership, public reaction to almost every type of broadcast had been tested, network broadcasting had been established and the economic basis upon which nationwide broadcasting now rests had been founded. A trail had been blazed that thereafter could be followed without hesitation.

In so far as radio broadcasting is concerned, this book is a significant chapter in communication history.

A Multichannel Microwave Radio Relay System.[2] H. S. Black, J. W. Beyer, T. J. Grieser, F. A. Polkinghorn. An 8-channel microwave relay system is described. Known to the Army and Navy as AN/TRC-6, the system uses radio frequencies approaching 5,000 megacycles. At these frequencies, there is a complete absence of static and most man-made interference. The waves are concentrated into a sharp beam and do not travel along the earth much beyond seeing distances. Other systems using the same frequencies can be operated in the near vicinity. The transmitter power is only one four-millionth as great as would be required with nondirectional antennas. The distance between sets is limited but by using intermediate repeaters communications are extended readily to longer distances. Short pulses of microwave power carry the intelligence of the eight messages utilizing *pulse position modulation* to modulate the

[1] Published by Harvard University Press, Cambridge, Massachusetts, 1946.
[2] *Elec. Engg., Trans. Sec.*, December 1946.

pulses and *time division* to multiplex the channels. The eight message circuits which each *AN/TRC*-6 system provides are high-grade telephone circuits and can be used for signaling, dialing, facsimile, picture transmission, or multichannel voice frequency telegraph. Two-way voice transmission over radio links totaling 1,600 miles, and one-way over 3,200 miles have been accomplished successfully in demonstrations.

Further Observations of the Angle of Arrival of Microwaves.[3] A. B. Crawford and William M. Sharpless. Microwave propagation measurements made in the summer of 1945 are described. This work, a continuation of the 1944 work reported elsewhere in this issue of the *Proceedings of the I.R.E. and Waves and Electrons,* was characterized by the use of an antenna with a beam width of 0.12 degree for angle-of-arrival measurements and by observations of multiple-path transmission.

The Effect of Non-Uniform Wall Distributions of Absorbing Material on the Acoustics of Rooms.[4] Herman Feshbach and Cyril M. Harris. The acoustics of rectangular rooms, whose walls have been covered by the non-uniform application of absorbing materials, is treated theoretically. Using appropriate Green's functions a general integral equation for the pressure distribution on the walls is derived. These equations show immediately that it is necessary to know *only* the pressure distribution on the treated surfaces to predict completely the acoustical properties of the room, such as the resonant frequencies, the decay constants, and the spatial pressure distribution. The integral equation is solved approximately using (1) perturbation method, and (2) approximate reduction of the integral equation to an equivalent transmission line. Criteria giving the range of validity of these approximations are derived. It was found useful to introduce a new concept, that of "*effective admittance,*" to express the results for the resonant frequency and absorption for then the amount of computation is reduced and the accuracy of the results is increased. The absorption of a patch of material was found as a function of the position of the absorbing material and was checked experimentally for a convenient case, an absorbing strip mounted on the otherwise hard walls of a rectangular room. Particular attention is given to the case where the acoustic material is applied in the form of strips. The results may then be expressed in series which converge very rapidly and are, therefore, amenable to numerical calculation. Approximate formulas are obtained which permit estimates of the diffusion of sound in a non-uniformly covered room. In agreement with experience, these equations show that diffusion increases with frequency and with the

[3] *Proc. I.R.E. and Waves and Electrons,* November 1946.
[4] *Jour. Acous. Soc. America,* October 1946.

number of nodes on the treated walls. The "interaction effect" of one strip on another is shown to decrease with an increase of the number of nodes. The results are then applied to the case of ducts with non-uniform distribution of absorbing material on its walls. Results are given which permit the calculation of the attenuation per unit length of duct. The methods of this paper hold for any distribution of absorbing material and also if the admittance is a function of angle of incidence.

High Current Electron Guns.[5] L. M. Field. This paper presents a survey of some of the problems and methods which arise in dealing with the design of high current and high current-density electron guns. A discussion of the general limitations on all electron gun designs is followed by discussion of single and multiple potential guns using electrostatic fields only. A further discussion of guns using combined electrostatic and magnetic fields and their limitations, advantages, and some possible design procedures follows.

Reflection of Sound Signals in the Troposphere.[6] G. W. Gilman, H. B. Coxhead, and F. H. Willis. Experiments directed toward the detection of non-homogeneities in the first few hundred feet of the atmosphere were carried out with a low power sonic "radar." The device has been named the *sodar*. Trains of audiofrequency sound waves were launched vertically upward from the ground, and echoes of sufficient magnitude to be displayed on an oscilloscope were found. Strong displays tended to accompany strong temperature inversions. During these periods, transmission on a microwave radio path along which the sodar was located tended to be disturbed by fading. In addition, relatively strong echoes were received when the atmosphere was in a state of considerable turbulence. There was a well-defined fine-weather diurnal characteristic. The strength of the echoes was such as to lead to the conclusion that a more complicated distribution of boundaries than those measured by ordinary meteorological methods is required in the physical picture of the lower troposphere.

A Cathode-Ray Tube for Viewing Continuous Patterns.[7] J. B. Johnson. A cathode-ray tube is described in which the screen of persistent phosphor is laid on a cylindrical portion of the glass. A stationary magnetic field bends the electron beam on to the screen, while rotation of the tube produces the time axis. When the beam is deflected and modulated, a continuous pattern may be viewed on the screen.

[5] *Rev. Mod. Phys.*, July 1946.
[6] *Jour. Acous. Soc. Amer.*, October 1946.
[7] *Jour. Applied Physics*, November 1946.

The Molecular Beam Magnetic Resonance Method. The Radiofrequency Spectra of Atoms and Molecules.[8] J. B. M. Kellogg and S. Millman. A new method known as the "Magnetic Resonance Method" which makes possible accurate spectroscopy in the low frequency range ordinarily known as the "radiofrequency" range was announced in 1938 by Rabi, Zacharias, Millman, and Kusch (R6, R5). This method reverses the ordinary procedures of spectroscopy and instead of analyzing the radiation emitted by atoms or molecules analyzes the energy changes produced by the radiation in the atomic system itself. Recognition of the energy changes is accomplished by means of a molecular beam apparatus. The experiment was first announced as a new method for the determination of nuclear magnetic moments, but it was immediately apparent that its scope was not limited to the measurement of these quantities only. It is the purpose of this article to summarize the more important of those successes which the method has to date achieved.

Metal-Lens Antennas.[9] Winston E. Kock. A new type of antenna is described which utilizes the optical properties of radio waves. It consists of a number of conducting plates of proper shape and spacing and is, in effect, a lens, the focusing action of which is due to the high phase velocity of a wave passing between the plates. Its field of usefulness extends from the very short waves up to wavelengths of perhaps five meters or more. The paper discusses the properties of this antenna, methods of construction, and applications.

Underwater Noise Due to Marine Life.[10] Donald P. Loye. The widespread use of underwater acoustical devices during the recent war made it necessary to obtain precise information concerning ambient noise conditions in the sea. Investigations of this subject soon led to the discovery that fish and other marine life, hitherto generally classified with the voiceless giraffe in noisemaking ability, have long been given credit for a virtue they by no means always practice. Certain species, most notably the croaker and the snapping-shrimp, are capable of producing noise which, in air, would compare favorably with that of a moderately busy boiler factory. This paper describes some of the experiments which traced these noises to their source and presents acoustical data on the character and magnitude of the disturbances.

Elastic, Piezoelectric, and Dielectric Properties of Sodium Chlorate and Sodium Bromate.[11] W. P. Mason. The elastic, piezoelectric, and di-

[8] *Rev. Mod. Phys.*, July 1946.
[9] *Proc. I.R.E. and Waves and Electrons*, November 1946.
[10] *Jour. Acous. Soc. America*, October 1946.
[11] *Phys. Rev.*, October 1 and 15, 1946.

electric constants of sodium chlorate ($NaClO_3$) and sodium bromate ($NaBrO_3$) have been measured over a wide temperature range. The value of the piezoelectric constant at room temperature is somewhat larger than that found by Pockels. The value of the Poisson's ratio was found to be positive and equal to 0.23 in contrast to Voigt's measured value of -0.51. At high temperatures the dielectric and piezoelectric constants increase and indicate the presence of a transformation point which occurs at a temperature slightly larger than the melting point. A large dipole piezoelectric constant (ratio of lattice distortion to dipole polarization) results for these crystals but the electromechanical coupling factor is small because the dipole polarization is small compared to the electronic and ionic polarization and little of the applied electrical energy goes into orienting the dipoles.

Paper Capacitors Containing Chlorinated Impregnants. Effects of Sulfur.[12] D. A. McLean, L. Egerton, and C. C. Houtz. Sulfur is an effective stabilizer for paper capacitors containing chlorinated aromatics, in the presence of both tin foil and aluminum foil electrodes. Sulfur has unique beneficial effects on power factor which are especially marked when tin foil electrodes are used. The value of R (Equation 4) can be used as an index of ionic conductivity in the impregnating compound. Diagnostic power factor measurements on impregnated paper are best made at low voltages. Electron diffraction studies give results in line with the previously published theory of stabilization. Several previous findings are reaffirmed: (a) the importance of all components of the capacitor in determining its initial properties and aging characteristics, (b) the superiority of kraft paper over linen, and (c) widely different behavior of capacitors employing different electrode metals.

A New Bridge Photo-Cell Employing a Photo-Conductive Effect in Silicon. Some Properties of High Purity Silicon.[13] G. K. Teal, J. R. Fisher, and A. W. Treptow. A pure photo-conductive effect was found in pyrolytically deposited and vaporized silicon films. An apparatus is described for making bridge type photo-cells by reaction of silicon tetrachloride and hydrogen gases at ceramic or quartz surfaces at high temperatures. The maximum photo-sensitivity occurs at 8400–8600A with considerable response in the visible region of the spectrum. The sensitivity of the cell appears about equivalent to that of the selenium bridge and its stability and speed of response are far better. For pyrolytic films on porcelain there are three distinct regions in the conductivity as a function of temperature. At low temperatures the electronic conductivity is given by the expression

[12] *Indus. & Engg. Chemistry*, November 1946.
[13] *Jour. Applied Physics*, November 1946.

$\sigma = Af(T)exp-(E/2kT)$. At temperatures between 227°C and a higher temperature of 400–500°C $\sigma = Aexp-(E/2kT)$, where E lies between 0.3 and 0.8 ev; and at high temperatures $\sigma = Aexp-(E/2kT)$, where $E = 1.12$ ev. The value 1.12 ev represents the separation of the conducting and non-conducting bands in silicon. The long wave limit of the optical absorption of silicon was found to lie at approximately 10,500A (1.18 ev). The data lead to the conclusion that the same electron bands are concerned in the photoelectric, optical, and thermal processes and that the low values of specific conductances found (1.8×10^{-5} ohm^{-1} cm^{-1}) are caused by the high purity of the silicon rather than by its polycrystalline structure.

Non-Uniform Transmission Lines and Reflection Coefficients.[14] L. R. WALKER and N. WAX. A first-order differential equation for the voltage reflection coefficient of a non-uniform line is obtained and it is shown how this equation may be used to calculate the resonant wave-lengths of tapered lines.

[14] *Jour. Applied Physics*, December 1946.

Contributors to this Issue

Harald T. Friis, E.E., Royal Technical College, Copenhagen, 1916; Sc.D., 1938; Assistant to Professor P. D. Pedersen, 1916; Technical Advisor at the Royal Gun Factory, Copenhagen, 1917–18; Fellow of the American Scandinavian Foundation, 1919; Columbia University, 1919. Western Electric Company, 1920–25; Bell Telephone Laboratories, 1925–. Formerly as Radio Research Engineer and since January 1946 as Director of Radio Research, Dr. Friis has long been engaged in work concerned with fundamental radio problems. He is a Fellow of the Institute of Radio Engineers.

Ray S. Hoyt, B.S. in Electrical Engineering, University of Wisconsin, 1905; Massachusetts Institute of Technology, 1906; M.S., Princeton, 1910. American Telephone and Telegraph Company, Engineering Department, 1906–07. Western Electric Company, Engineering Department, 1907–11. American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Hoyt has made contributions to the theory of loaded and non-loaded transmission lines and associated apparatus, theory of crosstalk and other interference, and probability theory with particular regard to applications in telephone transmission engineering.

W. D. Lewis, A.B. in Communication Engineering, Harvard College, 1935; Rhodes Scholar, Wadham College, Oxford; B.A. in Mathematics, Oxford, 1938; Ph.D. in Physics, Harvard, 1941. Bell Telephone Laboratories, 1941–. Dr. Lewis was engaged in radar antenna work in the Radio Research Department during the war; he is now engaged in microwave repeater systems research.

J. C. Lozier, A.B. in Physics, Columbia College, 1934; graduate physics student, Princeton University, 1934–35. R.C.A. Victor Manufacturing Company, 1935–36; Bell Telephone Laboratories, Inc., 1936–. Mr. Lozier has been engaged in transmission development work, chiefly on radio telephone terminals. During the war he was concerned primarily with the theory and design of servomechanisms.

VOLUME XXVI JULY, 1947 NO. 3

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Telephony by Pulse Code Modulation.... *W. M. Goodall* 395

Some Results on Cylindrical Cavity Resonators *J. P. Kinzer and I. G. Wilson* 410

Precision Measurement of Impedance Mismatches in Waveguide..... *Allen F. Pomeroy* 446

Reflex Oscillators..... *J. R. Pierce and W. G. Shepherd* 460

Abstracts of Technical Articles by Bell System Authors.. 682

Contributors to This Issue..... 691

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50¢ per copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORS

R. W. King J. O. Perrine

EDITORIAL BOARD

W. H. Harrison
O. B. Blackwell
H. S. Osborne
J. J. Pilliod
O. E. Buckley
M. J. Kelly
A. B. Clark
F. J. Feely

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are 50 cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1947
American Telephone and Telegraph Company

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXVI July, 1947 No. 3

Telephony By Pulse Code Modulation*

By W. M. Goodall

An experiment in transmitting speech by Pulse Code Modulation, or PCM, is described in this paper. Each sample amplitude of a pulse amplitude modulation or PAM signal is transmitted by a code group of ON-OFF pulses. 2^n amplitude values can be represented by an n digit binary number code. For a nominal 4 kc. speech band these n ON-OFF pulses are transmitted 8000 times a second. Experimental equipment for coding the PAM pulses at the transmitter and decoding the PCM pulses at the receiver is described. Experiments with this equipment indicate that a three-unit code appears to be necessary for a minimum grade of circuit, while a six- or seven-unit code will provide good quality.

Introduction

THIS paper describes an experiment in transmitting speech by PCM, or pulse code modulation. The writer is indebted to his colleagues in the Research Department, C. E. Shannon, J. R. Pierce and B. M. Oliver, for several interesting suggestions in connection with the basic principles of PCM given in this paper. Work on a different PCM system was carried on simultaneously in the Systems Development Department of the Bell Laboratories by H. S. Black. This in turn led to the development of an 8-channel portable system for a particular application. This system is being described in a forthcoming paper by H. S. Black and J. O. Edson.[1] A method for pulse code modulation is proposed in a U. S. Patent issued to A. H. Reeves.[2]

The material now presented is composed of three parts. The first deals with basic principles, the second describes the experimental PCM system, while the last discusses the results obtained.

Basic Principles

PCM involves the application of two basic concepts. These concepts are namely, the time-division principle and the amplitude quantization

* Paper presented in part at joint meeting of International Scientific Radio Union and Inst. Radio Engineers on May 5, 1947 at Washington, D. C.

[1] Paper presented on June 11, 1947 at A. I. E. E. Summer General Meeting, Montreal, Canada. Accepted for publication in forthcoming issue of *A. I. E. E. Transactions*.

[2] A. H. Reeves, *U. S. Patent* #2,272,070, Feb. 3, 1942, assigned to International Standard Electric Corp.; also, French patent #852,183, October 23, 1939.

principle. The essence of the time-division principle is that any input wave can be represented by a series of regularly occurring instantaneous samples, provided that the sampling rate is at least twice the highest frequency in the input wave.[3] For present purposes the amplitude quantization principle states that a complex wave can be approximated by a wave having a finite number of amplitude levels, each differing by one quantum, the size of the quantum jumps being determined by the degree of approximation desired.

Although other arrangements are possible, in this paper we will consider the application of these two basic principles in the following order. First the input wave is sampled on a time-division basis. Then each of the samples so obtained is represented by a quantized amplitude or integer number. Each of these integer numbers is represented as a binary number of n digits, the binary number system being chosen because it can readily be

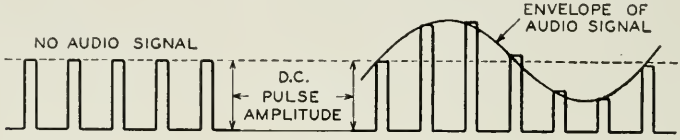


Fig. 1—Pulses in a PAM System.

represented by ON-OFF or two-position pulses. 2^n discrete levels can be represented by a binary number of n digits.[4] Thus, PCM represents each quantized amplitude of a time-division sampling process by a group of ON-OFF pulses, where these pulses represent the quantized amplitude in a binary number system.

The discussion so far has been in general terms. The principles just discussed will now be illustrated by examples.

Multiplex transmission of speech channels by sending short pulses selected sequentially from the respective speech channels, is now well known in the telephone art and is called time-division multiplex. When the pulses consist simply of short samples of the speech waves, their varying amplitudes directly represent the speech waves and the system is called pulse amplitude modulation or PAM. In PAM the instantaneous amplitude of the speech wave is sampled at regular intervals. The amplitude so obtained is trans-

[3] This is because the DC, fundamental and harmonics of the wave at the left in Fig. 1 all become modulated in the wave at the right, and if the highest modulating frequency exceeds half the sampling rate, the lower sideband of the fundamental will fall in the range of the modulating frequency and will not be excluded by the low-pass filter. The result is distortion.

[4] In a decimal system the digits can have any one of 10 values, 0 to 9 inclusive. In a binary system, the digits can have only two values, either 0 or 1.

mitted as a pulse of corresponding amplitude. In order to transmit both positive and negative values a constant or d-c value of pulse amplitude can be added. (See Fig. 1.) When this is done positive values of the information wave correspond to pulse amplitudes greater than the constant value while negative values correspond to pulse amplitudes less than the constant value. At the receiver a reproduction of the original speech wave will be obtained at the output of a low-pass filter.

The PCM system considered in this paper starts with a PAM system and adds equipment at the terminals to enable the transmission of a group of ON-OFF pulses or binary digits to represent each instantaneous pulse amplitude of the PAM system. Representation of the amplitude of a single PAM pulse by a finite group of ON-OFF pulses or binary digits requires quantization of the audio wave. In other words, we cannot represent the actual amplitude closer than $\frac{1}{2}$ "quantum". The number of amplitude levels required depends upon the grade of circuit desired. The disturbance which results from the quantization process has been termed quantizing noise. For this type of noise a signal-to-noise ratio of 33 db would be obtained for 32 amplitude levels and this grade of circuit was deemed sufficiently good for a preliminary study. These 32 amplitude levels can be obtained with 5 binary digits, since $32 = 2^5$.

Figure 2 shows how several values of PAM pulse amplitude can be represented by this binary code. The first column gives the digit pulses which are sent between the transmitter and receiver while the second column shows the same pulse pattern with each pulse weighted according to its assigned value, and the final column shows the sum of the weighted values. The sum, of course, represents the PAM pulse to the nearest lower amplitude unit. The top row where all the digits are present shows, in the middle wave form, the weighted equivalent of each digit pulse. By taking different combinations of the five digits all integer amplitudes between 31 and 0 can be represented. The examples shown are for 31, 18, 3, and 0.

Referring to Fig. 3 sampling of the audio wave (a) yields the PAM wave (b). The PAM pulses are coded to produce the code groups or PCM signal (c). The PCM pulses are the ones sent over the transmission medium. For a sampling rate of 8000 per second, there would be 8000 PAM pulses per second for a single channel. The digit pulse rate would be 40,000 pps for a five-digit code. For a time-division multiplex of N channels both of these pulse rates would be multiplied by N.

Wave form (d) shows the decoded PAM pulses where the amplitudes are shown under the pulses. The original audio wave is repeated as wave form (e). It will be noted that the received signal is delayed by one PAM pulse interval. It is also seen that the decoded pulses do not fit exactly on this curve. This is the result of quantization and the output of the low-pass

filter will contain a quantizing disturbance not shown in (e) which was not present in the input signal.

A signal that uses regularly occurring ON-OFF pulses can be "regenerated" and repeated indefinitely without degradation. A pulse can be "regenerated" by equipment which transmits an undistorted pulse provided a somewhat distorted pulse is received, and transmits nothing otherwise.

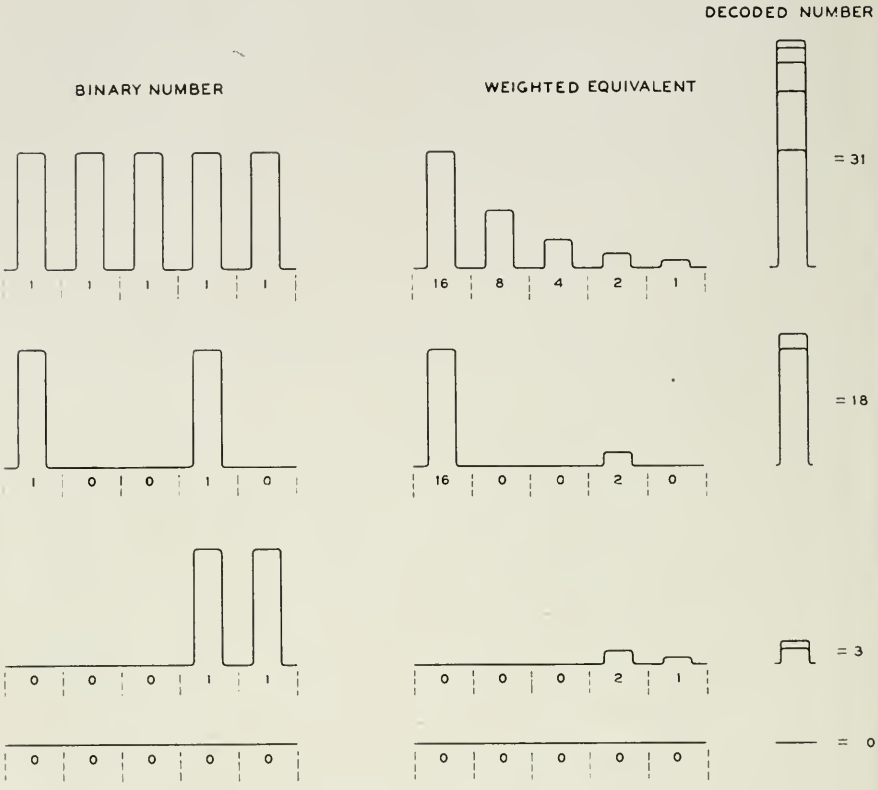


Fig. 2—Binary and decimal equivalents.

Thus, the received signal at the output of the final decoder is of the same quality as one produced by a local monitoring decoder. To accomplish this result, it is necessary, of course, to regenerate the digit pulses before they have been too badly mutilated by noise or distortion in the transmission medium.

The regenerative property of a quantized signal can be of great importance in a long repeated system. For example, with a conventional system each repeater link of a 100-link system must have a signal-to-noise ratio 20 db

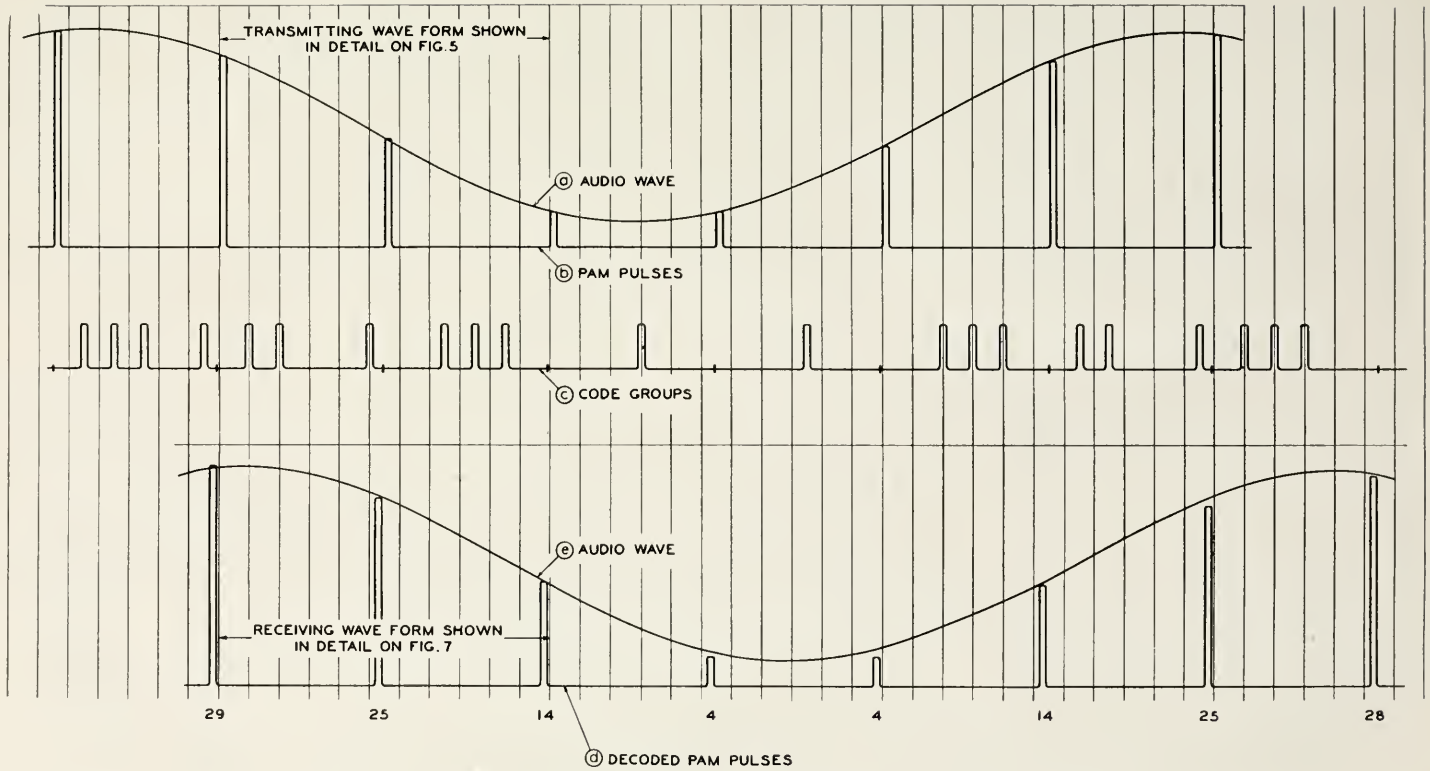


Fig. 3—PAM and PCM transmitting and receiving wave forms (amplitude vs. time).

better than the complete system. For PCM, however, with regenerative repeaters the required signal-to-noise ratio in the radio part of the system is independent of the number of links. Hence, we have a method of transmission that is ideally suited to long repeated systems.

At this point we might consider the bandwidth required to send this type of signal. For a 5-digit code the required band is somewhat less than 5 times that required for a PAM system. It is somewhat less than 5 times because in a multiplex system crosstalk becomes a serious problem. In a PAM system this crosstalk would add up on a long system in somewhat the same manner as noise. In order to reduce the crosstalk it would probably be necessary to use a wider band for the PAM repeater system than would be required for a single-link system. For PCM, on the other hand, by using regeneration the whole system requirement for crosstalk can be used for each link. In addition, a relatively greater amount of crosstalk can be tolerated since only the presence or absence of a pulse needs to be determined. Both of these factors favor PCM. This is a big subject and for the present we need only conclude that from considerations of the type just given the bandwidth penalty of PCM is not nearly as great as might first be expected.

The same two factors that were mentioned in connection with crosstalk also apply to noise, and a PCM signal can be transmitted over a circuit which has a much lower signal-to-noise ratio than would be required to transmit a PAM signal, for example.

Hence, we conclude that PCM for a long repeated system has some powerful arguments on its side because of its superior performance even though it may require somewhat greater bandwidth. There are other factors where PCM differs from more conventional systems but a discussion of these factors is beyond the scope of this paper.

The previous discussion may be summarized as follows: One begins with a pulse amplitude modulation system in which the pulse amplitude is modulated above and below a mean or d-c value as indicated in Fig. 1. It is assumed that it will be satisfactory to limit the amplitude range to be transmitted to a definite number of amplitude levels. This enables each PAM pulse to be represented by a code group of ON-OFF pulses, where the number of amplitude levels is given by 2^n, n being the number of elements in each code group. With this system the digit pulses can be "regenerated" and the quality of the overall transmission system can be made to depend upon the terminal equipment alone.

Experimental PCM Equipment

The experimental coder used in these studies might be designated as one of the "feedback subtraction type". It functions as follows: Each PAM pulse is stored as a charge on a condenser in a storage circuit. (See Fig. 4.)

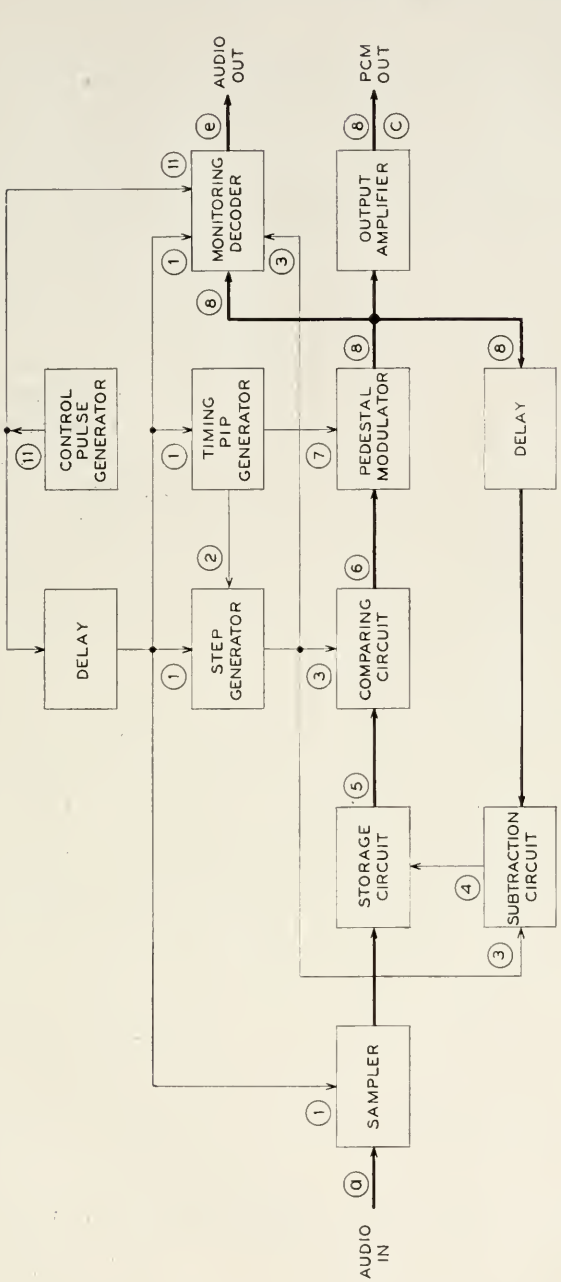


Fig. 4—PCM Transmitter block schematic.

The voltage across this condenser is compared with a reference voltage. The magnitude of this reference voltage corresponds to the d-c pulse amplitude of Fig. 1. The voltage has a magnitude of 16 units. If the magnitude of the condenser voltage exceeds the magnitude of the 16-unit voltage, a positive pedestal voltage is obtained in the output of the comparing circuit. This pedestal voltage is amplified, limited and applied to the pedestal modulator. The pedestal modulator serves as a gate for timing pulses from the timing pip generator. If the pedestal voltage and timing pulse are applied simultaneously to the pedestal modulator, a pulse is obtained in the output. In the present case this pulse corresponds to the presence of the 16-unit digit in the code group which represents this PAM pulse. This digit pulse after amplification and limiting is (1) sent out over the line (PCM out) and (2) fed back through a suitable delay circuit to a subtraction circuit. The function of the subtraction circuit is to subtract a charge from the condenser corresponding to the 16-unit digit. The charge remaining on the condenser is now compared with a new reference voltage which is $\frac{1}{2}$ the magnitude of the first reference voltage or 8 units. If the magnitude of the voltage across the condenser exceeds this new reference voltage the above process is repeated and the second digit pulse is transmitted and another charge, this time corresponding to the 8-unit digit, is subtracted from the remaining charge upon the condenser.

If the magnitude of the voltage across the condenser is less than the reference voltage, in either case above, then no pedestal will be produced and no digit pulse be transmitted. Since no pulse is transmitted, no charge will be subtracted from the condenser. Thus the charge remaining upon the condenser after each operation represents the part of the original PAM pulse remaining to be coded. The reference voltage wave consists of a series of voltages each of which is $\frac{1}{2}$ of the preceeding one. There is one step on the reference voltage function for each digit to be coded.

A better understanding of the coding process can be had by reference to the various wave forms involved. For completeness, wave forms from audio input to the coded pulse signal are shown for the transmitter in Figs. 3 and 5 and from the coded pulse signal to audio output for the receiver in Figs. 7 and 3. In the diagram the abscissas are time and the ordinates are amplitudes. Some of these wave forms have already been discussed in connection with Fig. 3. Since the coder functions in the same manner for each PAM pulse the detailed wave forms of the coding and decoding processes are shown for only two amplitudes. The block schematic for the transmitter is given on Fig. 4, while that for the receiver is given in Fig. 6. The letters on Figs. 4 and 6 refer to the wave forms on Fig. 3, while the numbers refer to the wave forms in Figs. 5 and 7.

Fig. 5—Detailed wave forms for PCM Transmitter (amplitude vs. time).

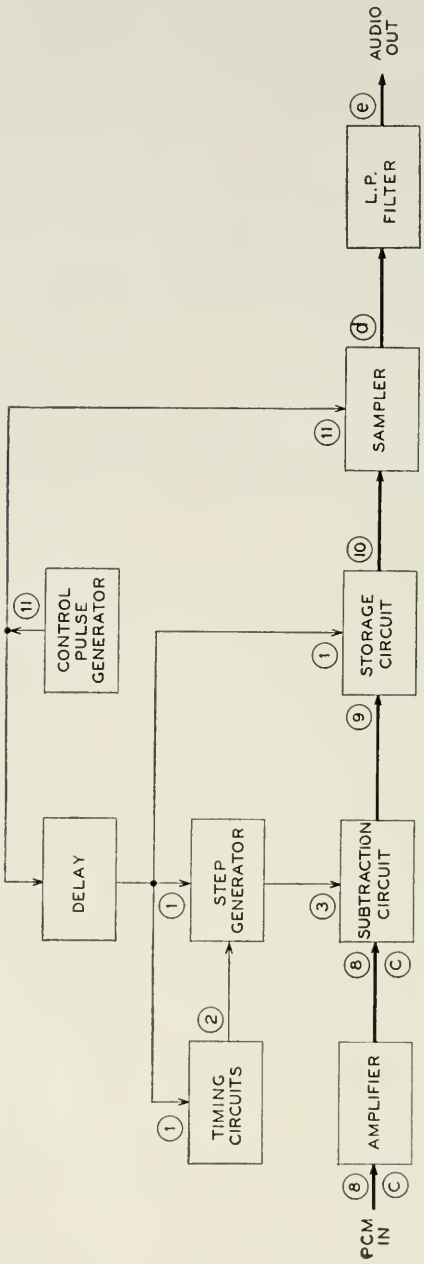


Fig. 6—PCM Receiver block schematic.

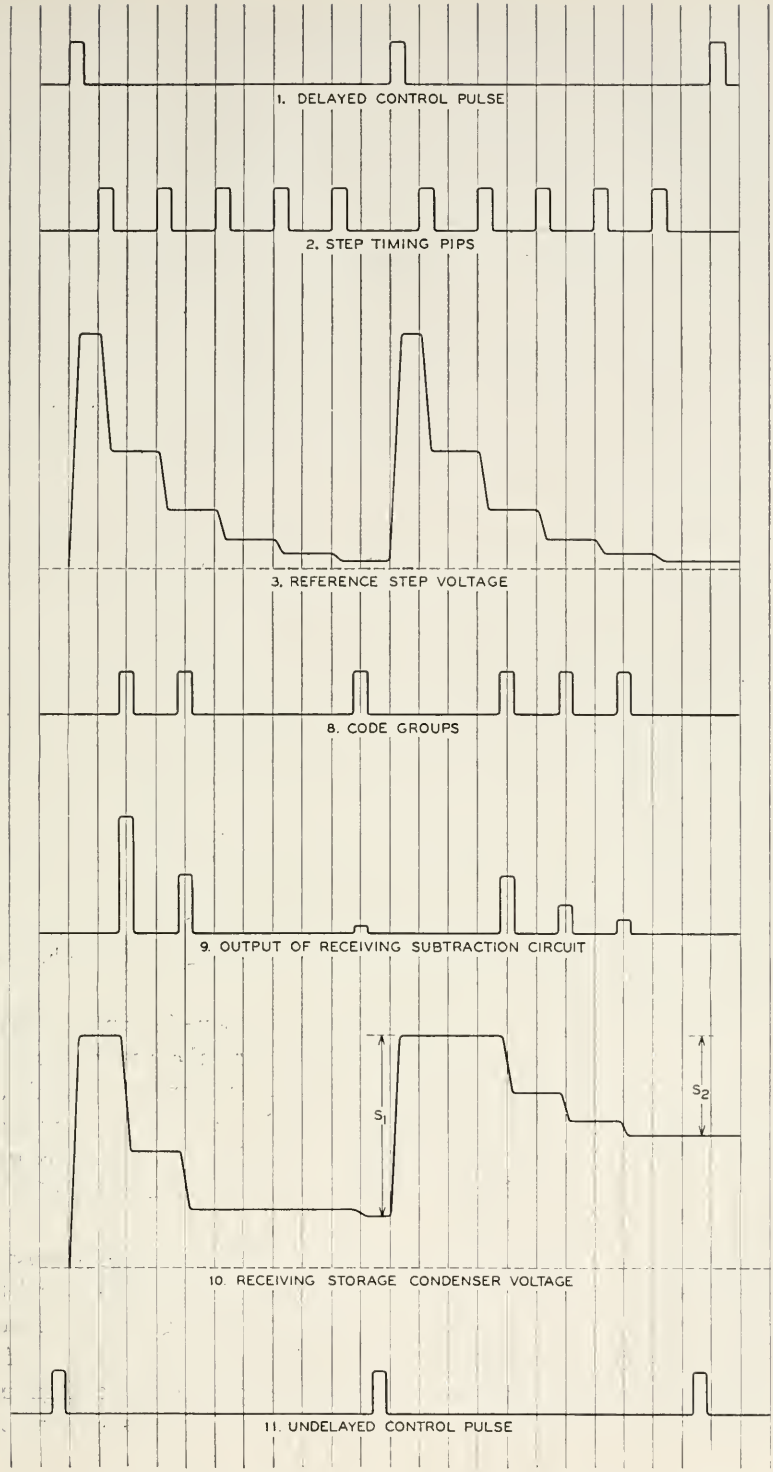


Fig. 7—Detailed wave forms for PCM Receiver (amplitude vs. time).

Referring to Figs. 4 and 5, the "delayed control pulse" Curve 1 is the principal timing pulse for the transmitting coder. It is used to sample the audio wave and to start the step and timing-pip generators. Two sets of timing-pips are produced; one, Curve 2, is used to generate the reference step voltage while the other, Curve 7, is used for timing the digit pulses. The reference step voltage, Curve 3, is used in the comparing circuit and in the subtraction circuit. Curve 4 gives the output of the subtraction circuit, while Curve 5 is the voltage on the storage condenser. The next plot gives Curves 3 and 5 superimposed; the shaded area on this plot corresponds to the time during which a pedestal voltage is generated. The pedestal voltage is given by Curve 6, and the output of the pedestal modulator is given by Curve 8. This last curve is a plot of the two code groups corresponding to the two PAM pulses being coded.

In studying these wave forms it will be noted that the delayed control pulse, the two sets of timing-pips and the reference step voltage curves are the same for each code group. On the other hand the storage condenser voltage, the pedestal voltage, the group of code pulses, and the group of pulses from the subtraction circuit are different for each code group.

It will be recalled that a pedestal voltage is produced during the time that the condenser voltage exceeds the reference step voltage. The leading edge of each pedestal pulse is generated by the falling part of the reference step voltage. The trailing edge of each pedestal pulse is produced by the falling part of the condenser voltage. This drop in condenser voltage is the result of the operation of the subtraction circuit. The output of the subtraction circuit depends upon the delayed digit pulse which has just been passed by the pedestal pulse. Its magnitude depends upon the reference voltage step that applies to the particular digit being transmitted. The function of the delay in the feedback path is to allow the outgoing digit pulse to be completed before the pedestal is terminated.

It is seen that the pedestal voltage contains the same information as the transmitted code groups. Under ideal conditions the use of auxiliary timing pulses would not be required. However, in a practical circuit the leading edge of the pedestal varies, both as to relative timing and as to rate of rise. Under these conditions the auxiliary timing-pips permit accurate timing of the outgoing PCM pulses, as well as constant pulse shape for the input to the subtraction circuit.

Summarizing the foregoing it is seen that in the coder under discussion a comparison is made for each digit between a reference voltage and the voltage across a storage condenser. Initially the voltage across this condenser represents the magnitude of the PAM pulse being coded. After each digit the voltage remaining on the condenser represents the magnitude of the original PAM pulse remaining to be coded. A pedestal voltage is

obtained in the output of the comparing circuit whenever the storage condenser voltage exceeds the reference step voltage.
This pedestal, if present, allows a timing pulse to be sent out as a digit of the code group. This digit pulse is also delayed and fed back to a subtraction circuit which reduces the charge on the condenser by a magnitude corresponding to the digit pulse just transmitted. This process is repeated step by step until the code is completed.

Synchronizing the two control pulse generators, one at the transmitter and one at the receiver, is essential to the proper operation of the equipment. This may be accomplished in a variety of ways. The best method of synchronizing to use would depend upon the application. Although the control could easily be obtained by transmitting a synchronizing pulse over the line, the equipment would have been somewhat more complicated and for these tests a separate channel was used to synchronize the control pulse generators at the terminals.

Having thus established the timing of the receiving control pulse generator shown in Fig. 6 relative to the received code groups, the receiver generates a new set of waves as shown in Fig. 7. Except for delay in the transmission medium, the first three curves are the same as those shown in Fig. 5 for the transmitter. (1) is the delayed control pulse, (2) is the step timing wave, and (3) is the reference step voltage. Curve 8 is the received code group and (9) is the output current of the subtraction circuit. (10) gives the wave form of the voltage across the receiving storage circuit, and (11) gives the curve for the undelayed control pulse.

The receiver functions as follows: The storage condenser is charged to a fixed voltage by each delayed control pulse. The charge on the condenser is reduced by the output of the subtraction circuit. The amount of charge that is subtracted depends upon which digit of the group produces the subtraction pulse. This amount is measured by the reference step voltage. At the end of the code group the voltage remaining on the condenser is sampled by the undelayed control pulse.

It is seen that the storage subtraction circuits in the transmitter and receiver function in similar ways. In the transmitter the original voltage on the condenser depends upon the audio signal, and after the coding process this voltage is substantially zero. The receiver starts with a fixed maximum voltage and after the decoding process the sample that is delivered to the output low-pass filter is given by the voltage reduction of the condenser during the decoding process. Except that the conditions at beginning and end of the coding and decoding periods are different as discussed above, the subtraction process is the same for both units.

The monitoring decoder in the transmitter operates in the same manner described above, except that it employs the various waves already generated for other uses in the transmitter (see Fig. 4).

Experimental Results

An experimental system was set up as shown in Fig. 8. The pulse code modulator, radio transmitter, and antenna comprised the transmitting terminal; while an antenna, radio receiver and pulse code demodulator were used for the receiving terminal. A short air-path separated the terminals. The transmitter used a pulsed magnetron oscillator and the receiver employed a broad-band superheterodyne circuit. The results obtained with this system were similar to those obtained by connecting the pulse code

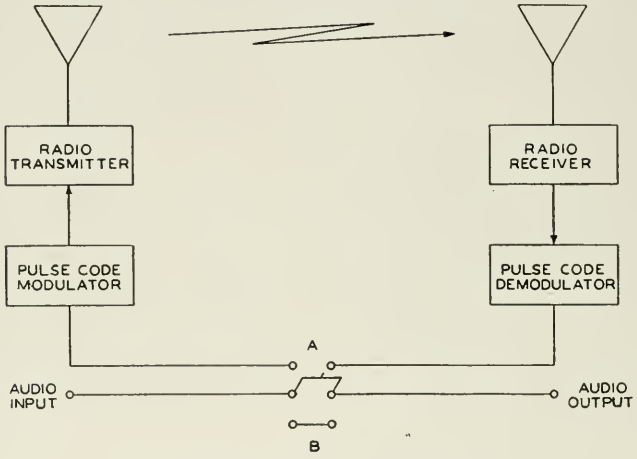


Fig. 8—Block diagram of PCM system.

modulator and demodulator together without the radio equipment. In fact, unless a large amount of attenuation was inserted in the path the presence of the radio circuit could not be detected.

It was possible to adjust the PCM transmitter so that different numbers of digits could be produced. A brief study was made of the number of digits required. It was found that, with regulated volume, a minimum of three or four digits was necessary for good intelligibility for speech though, surprisingly enough, a degree of intelligibility was obtained with a single one. With six digits both speech and music were of good quality when regulated volume was used. Even with six digits, however, it was possible to detect the difference between PCM and direct transmission in A-B tests. This could be done most easily by a comparison of the noise in the two systems. If unregulated volume were used several more digits would probably be desirable for high quality transmission.

In listening to the speech transmitted over the PCM system one obtained the impression that the particular sound patterns of a syllable or a word

could be transmitted with three or four digits. If the volume range of the talker varied it would be necessary to add more digits to allow for this variation. Over and above these effects, however, the background noise which is present to a greater or lesser extent in all communication circuits, is quantized by the PCM system. If the size of the quanta or amplitude step is too large the circuit will have a characteristic sound, which can easily be identified. Since the size of the quanta is determined by the number of digits, it is seen that the number of digits required depends not alone upon the speech but also upon the background noise present in the input signal.

Summarizing, experimental results obtained indicate that at least 3 digits are desirable for a minimum grade of circuit and that as many as 6 or more will provide for a good quality circuit. If we wish to transmit a nominal speech band of 4000 cycles, PCM requires the 8000 pulses per second needed by any time-division system, multiplied by the number of digits transmitted. The extra bandwidth required for PCM however, buys some real advantages including freedom from noise, crosstalk and signal mutilation, and ability to extend the circuit through the use of the regenerative principle.

The writer wishes to acknowledge the assistance of Mr. A. F. Dietrich in the construction and testing of the PCM equipment discussed in this paper.

Some Results on Cylindrical Cavity Resonators

By J. P. KINZER and I. G. WILSON

Certain hitherto unpublished theoretical results on cylindrical cavity resonators are derived. These are: an approximation formula for the total number of resonances in a circular cylinder; conditions to yield the minimum volume circular cylinder for an assigned Q; limitation of the frequency range of a tunable circular cylinder as set by ambiguity; resonant frequencies of the elliptic cylinder; resonant frequencies and Q of a coaxial resonator in its higher modes; and a brief discussion of fins in a circular cylinder.

The essential results are condensed in a number of new tables and graphs.

Introduction

THE subject of wave guides and the closely allied cavity resonators was of considerable interest even prior to 1942, as shown in the bibliography. It is believed that this bibliography includes virtually everything published up to the end of 1942. During the war, many applications of cavity resonators were made. Among these was the use of a tunable circular cylinder cavity in the $TE\ 01n$ mode as a radar test set; this has been treated in previous papers.[1,2] During this development, a number of new theoretical results were obtained; some of these have been published.[2] Here we give the derivation of these results together with a number of others not previously disclosed.

In the interests of brevity, an effort has been made to eliminate all material already published. For this reason, the topics are rather disconnected, and it is also assumed that the reader has an adequate background in the subject, such as may be obtained from a study of references 3 to 7 of the bibliography, or a text such as Sarbacher and Edson.[8]

A convenient reference and starting point is afforded by Fig. 1, taken from the Wilson, Schramm, Kinzer paper.[2] This figure also explains most of the notation used herein.

Acknowledgement

In this work, as in any cooperative scientific development, assistance and advice were received from many individuals and appropriate appreciation therefor is herewith extended. In some cases, explicit credit for special contributions has been given.

Contents

1. Approximation formula for number of resonances in a circular cylindrical cavity resonator.
2. Conditions for minimum volume for an assigned Q.

NORMAL WAVELENGTHS
$\lambda = \frac{2}{\sqrt{\left(\frac{l}{a}\right)^2 + \left(\frac{m}{b}\right)^2 + \left(\frac{n}{L}\right)^2}}$
SAME AS TM MODES
$\lambda = \frac{2}{\sqrt{\left(\frac{2r_{\ell m}}{\pi a}\right)^2 + \left(\frac{n}{L}\right)^2}}$ $\left(fa\right)^2 = \left(\frac{cr_{\ell m}}{\pi}\right)^2 + \left(\frac{cn}{2}\right)^2 \left(\frac{a}{L}\right)^2$ $c = \sqrt{\mu\epsilon}$ = VELOCITY OF ELECTROMAGNETIC WAVES IN DIELECTRIC   f = FREQUENCY
SAME FORM AS FOR CYLINDER $r_{\ell m}$ HAS DIFFERENT VALUES

)1*n* mode

xial reso-

y are ob-

(1)

The dis-
t, can be

quency f_0
e the true
es being a
ped curve
l approxi-

inces of a
Bolt[9] and
to apply

e from the

sonant fre-

represent-
to find the

$$R = \frac{\pi a f_0}{c}.$$

de shape factors for recta

Cc
ators
of re
cular
circu
resor
discu
Th

THI
of
It is beli
up to th
nators w
cavity ir
vious pa
results w
the deriv
ously dis
In the
material
nected, a
in the su
of the bi
A conv
the Wils
of the nc

In this
advice we
therefor i
contributi

1. Appr
drica
2. Conc

3. Limitation of frequency range of a tunable cavity in the $TE\,01n$ mode as set by ambiguity.
4. Resonant frequencies of an elliptic cylinder.
5. Resonant frequencies and Q of higher order modes of a coaxial resonator.
6. Fins in a circular cylinder.

Approximation Formula for Number of Resonances in a Circular Cylinder

From Fig. 1, the resonant frequencies of the cylindrical cavity are obtained from the equation:

$$(fa)^2 = \left(\frac{cr}{\pi}\right)^2 + \left(\frac{cna}{2L}\right)^2 \tag{1}$$

in which r is written in place of $r_{\ell m}$, to simplify the equations. The distribution of the resonant frequencies, starting with the lowest, can be approximated by a continuous function

$$N \approx F(f_0) \approx G(\lambda_0)$$

where N represents the total number of resonances up to a frequency f_0 or a wavelength λ_0. This is bound to be an approximation, since the true function F is discontinuous (or stepped) by virtue of the resonances being a series of discrete values. For practical purposes, if F fits the stepped curve so that the steps fluctuate above and below F, it will be a useful approximation.

Derivation of such a formula as applied to the acoustic resonances of a rectangular box has recently been a subject of investigation by Bolt[9] and Maa.[10] Only slight modifications of their method need be made to apply to the present situation.

Multiply (1) thru by $\left(\frac{\pi}{c}\right)^2$:

$$\left(\frac{\pi af}{c}\right)^2 = r^2 + \left(\frac{\pi an}{2L}\right)^2.$$

Hence, if a point $\left(r, \frac{\pi an}{2L}\right)$ is plotted on the XY plane the distance from the origin to this point will be $\frac{\pi af}{c}$ and hence a measure of the resonant frequency. If all such points are plotted, they will form a lattice representing all the possible modes of resonance. The problem, then, is to find the number of lattice points in a quadrant of a circle with radius, $R = \frac{\pi af_0}{c}$.

The values of the Bessel zero, r, are not evenly spaced along the X axis; indeed the density, or number per unit distance, increases as r increases. Let the density be $p(x)$. Then the problem becomes one of finding the weight of a quadrant of material whose density varies as $p(x)$.

Suppose the expression for M, the number of zeros r, less than some value x, is of the form

$$M = Ax^2 + Bx$$

whence, by differentiation,

$$p(x) = 2Ax + B.$$

The weight, W, of the quadrant of a circle of radius R is then, by integration,

$$W = \frac{2}{3} AR^3 + \frac{\pi}{4} BR^2.$$

Since there are $\frac{2L}{\pi a}$ lattice points per unit distance along the Y axis, $\frac{2LW}{\pi a}$ is apparently the total number of points in the quadrant. However, there are two small corrections to consider. First is that in this procedure a lattice point is represented by an area and for the points along the X axis half the area, i.e., a strip $\frac{\pi a}{4L}$ wide lying in the adjacent quadrant, has been omitted. Second is that the restriction $n > 0$ for TE modes eliminates half the points along the X axis. As it happens, these corrections just cancel each other. Thus we have

$$N = \frac{16\pi A}{3} \frac{V}{\lambda_0^3} + \frac{\pi B}{2} \frac{S}{\lambda_0^2}$$

in which

$$V = \frac{\pi a^2 L}{4} \qquad S = \pi aL \qquad \lambda_0 = \frac{c}{f_0}.$$

From a tabulation[11] of the first 180 values of r, the empirical values $A = 0.262$, $B = 0$ were obtained. This gives

$$N = 4.39 \frac{V}{\lambda_0^3}.$$

Subsequently, from an analysis of over a thousand modes in a "square cylinder" ($a = L$), Dr. Alfredo Baños, formerly of M.I.T. Radiation Laboratory, has calculated the empirical formula

$$N = 4.38 \frac{V}{\lambda_0^3} + 0.089 \frac{S}{\lambda_0^2} \tag{2}$$

from which $A = 0.262$, $B = 0.057$. These values give better agreement with the 180 tabulated values of r.

There is a two-fold degeneracy in a circular cylinder for modes with $\ell > 0$, which is removed, for example, when the cylinder is made elliptical. The total number of modes, then, counting degeneracies twice, is about $2N$, which brings (2) in line with the general result that, in any cavity resonator, the total number of modes is of the order $\frac{8\pi}{3}\frac{V}{\lambda_0^3}$.

Minimum Volume of Circular Cylinder for Assigned Q

In practical applications of resonant cavities, the conditions of operation may require high values of Q which can be attained only by the use of high order modes. The total number of modes, most of which are undesired, can then be reduced only by making the cavity volume as small as possible, consistent with meeting the requirement on Q.

It will be shown that, for a cylinder, operation in the $TE\ 01n$ mode very probably gives the smallest volume for an assigned Q.

Statement of Problem

When the relative proportions (the shape) of a cavity and the mode of oscillation are fixed, both the Q and the volume, V, of the cavity are functions of the operating wavelength, λ. Since we are primarily interested in the relationship between Q and V, with λ fixed, some simplification can be made by eliminating λ as a parameter. This may be done by a change of variables to $Q\frac{\delta}{\lambda}$ and $\frac{V}{\lambda^3}$, respectively; to simplify the typography, these quantities will be denoted by single symbols:

$$P \equiv Q\frac{\delta}{\lambda}$$

$$W \equiv \frac{V}{\lambda^3}.$$

We are, consequently, interested in the following specific problem:

In a circular cylindrical resonator, which is the optimum mode family and what is the corresponding shape to obtain the smallest value of W for a preassigned value of P?

A rigorous solution cannot be obtained by the methods of elementary calculus, since P is not a continuous function of the mode of oscillation. However, a possible procedure is to assume continuity, and examine the relation between P and W under this assumption. If sufficiently positive results are obtained, the conclusions may then be carried over to the discontinuous (i.e., the physical) case with reasonable assurance that, except

perhaps for special values, the correct answer is obtained. We proceed on this basis.

Solution

To permit a more coherent presentation of the arguments, only their general outline follows. More mathematical details are given later.

We start with the formulas for $Q\frac{\delta}{\lambda}$ $(= P)$ as given in Fig. 1.

The first operation is to show that, under comparable conditions, i.e., λ, r, n fixed, the TE $0mn$ modes give the highest values of P. That this is plausible can be seen in a general manner from the equations as they stand. For the TE modes, if $\ell = 0$, the numerator of the fraction is largest. Also, P simplifies, and the denominator roughly reduces the expression in square brackets to the 1/2 power. Now compare this expression with those for the TM modes. That for the TM modes ($n > 0$) is smaller because of the factor $(1 + R)$ in the denominator. Finally, that for the TM modes ($n = 0$) is still smaller, because $1 < (1 + p^2R^2)^{1/2}$.

This leaves only the TE $0mn$ modes to be considered, and the next step is to show that $m = 1$ is the most favorable value. Since the relation between P and W is complicated, a parameter φ is introduced, with φ defined by

$$\tan \varphi = pR. \tag{3}$$

The resulting parametric equations are:

$$P = \frac{r}{2\pi} \frac{1}{\cos^3 \varphi + \frac{1}{p} \sin^3 \varphi} \tag{4}$$

$$W = \frac{pr^3}{4\pi^2} \frac{1}{\cos^2 \varphi \sin \varphi}. \tag{5}$$

For each of the discrete values of r and n (n is related to p) then, plots of P vs W can be prepared as shown in Fig. 2 for the TE $01n$ modes.

Inspection of Fig. 2 shows that the best value of Q does not correspond to a minimum of W or a maximum of P for a given value of n, but rather to a point on the "envelope" of the curves. To get the envelope, we assume p to be continuous and proceed in the standard manner. It turns out that, by solving (4) for p in terms of P, r and φ, substituting the resulting expression in W, and setting $\frac{\partial W}{\partial \varphi} = 0$ an equation is obtained which, when solved for φ, gives the values of φ which lie on the envelope.

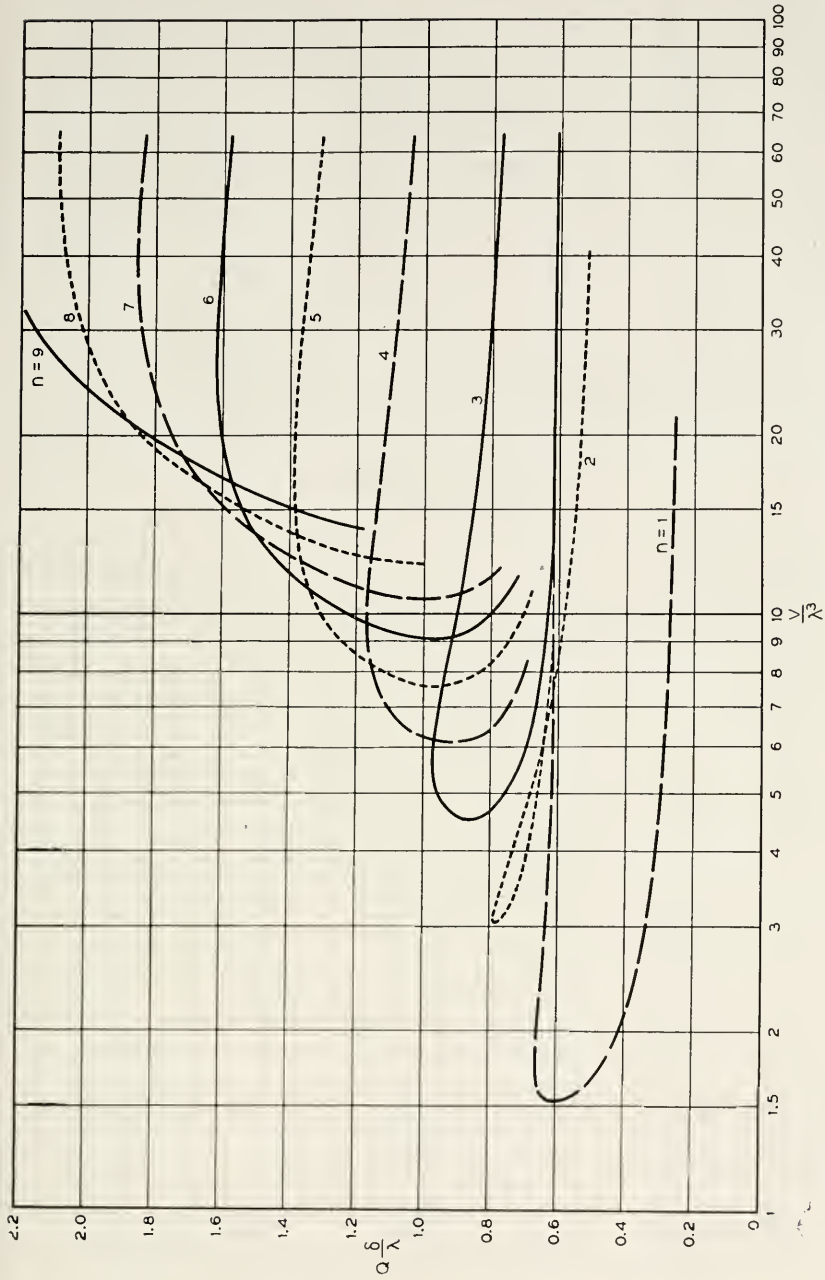


Fig. 2—Relation between $Q\frac{\delta}{\lambda}$ (= P) and $\frac{V}{\lambda^3}$ (= W) in cylindrical cavity resonator operating in $TE\ 01n$ mode.

We next substitute this expression for φ in W and calculate $\frac{\partial W}{\partial r}$ assuming now that r is continuous, and find that W has no minimum. Practically, this means that the smallest value of r should be used, i.e., the $TE\,01n$ mode.

Finally, since from Fig. 2 it is seen that the envelope is reasonably smooth for values of $Q\frac{\delta}{\lambda} > 1$, the expression for φ derived on the assumption of continuous p is used to obtain a simple relation of great utility in practical cavity design.

Details of solution

In (3), since R must be finite for a physical cylinder, $0 < \tan\varphi < \infty$, $0 < \sin\varphi < 1$, and $0 < \cos\varphi < 1$. Hence we may always divide by $\sin\varphi$ or $\cos\varphi$. Note that φ ranges between 0° and 90°.

From Fig. 1,

$$k = \frac{2r(1 + p^2R^2)^{1/2}}{a}$$

whence

$$k\sin\varphi = \frac{2prR}{a} \tag{6}$$

$$k\cos\varphi = \frac{2r}{a}. \tag{7}$$

We define W by:

$$W \equiv \frac{V}{\lambda^3} = \frac{\pi a^3}{4R}\,\frac{k^3}{8\pi^3}. \tag{8}$$

Substituting (6) and (7) in (8),

$$W = \frac{pr^3}{4\pi^2}\,\frac{1}{\cos^2\varphi\sin\varphi}. \tag{5'}$$

Substitution of (3) into the expression for $Q\frac{\delta}{\lambda}$ $(= P)$ for the TE modes as given in Fig. 1 yields, after some manipulation

$$P = \frac{r}{2\pi}\cdot\frac{1 - (\ell/r)^2}{\cos^3\varphi + \frac{1}{p}\sin^3\varphi + \left(\cos\varphi - \frac{1}{p}\sin\varphi\right)(\ell/r)^2\sin^2\varphi}.$$

To show that any value of $\ell > 0$ reduces P below its value when $\ell = 0$, let

$$a = \cos^3\varphi + \frac{1}{p}\sin^3\varphi$$

$$b = \left(\cos\varphi - \frac{1}{p}\sin\varphi\right)\sin^2\varphi$$

$$c = (\ell/r)^2.$$

It suffices to show that

$$\frac{1}{a} > \frac{1-c}{a+bc}$$

where the question is in doubt because b may take on negative values. If the inequality is to be valid, it is necessary only that $(b + a) > 0$, that is, $\cos\varphi > 0$. Hence, for the TE modes, only $\ell = 0$ needs be considered. For this case, the expression for P simplifies to

$$P = \frac{r}{2\pi}\frac{1}{\cos^3\varphi + \frac{1}{p}\sin^3\varphi}. \qquad (4')$$

For the TM modes, there is similarly obtained

$$P = \frac{r}{2\pi}\frac{1}{\cos\varphi + \frac{1}{p}\sin\varphi} \qquad n > 0 \qquad (9)$$

$$P = \frac{r}{2\pi}\frac{\cos\varphi}{\cos\varphi + \frac{1}{2p}\sin\varphi} \qquad n = 0. \qquad (10)$$

It is easy to show, since $\cos\varphi < 1$ and $\sin\varphi < 1$, that both (9) and (10) are less than (4').

Hence we have shown that, under comparable conditions, i.e., r and p constant, the TE $0mn$ modes have higher values of P than any others. There is one flaw in the argument, viz., r takes on discrete values and cannot be made the same for all modes. It is conceivable, therefore, that for some specific values of P, a mode other than the TE $0mn$ can be found which gives a smaller W than either of the two "adjacent" TE $0mn$ modes, one having a value of r higher, the other lower, than the supposed high-P mode. This situation requires further refinement, and hence complication, in the analysis; we pass over this point.

Having so far indicated that the TE $0mn$ modes are the best, our next objective is find the best value of m, if possible.

By use of the parametric equations (4) and (5), Fig. 2 has been plotted for $r = 3.83$ ($TE\ 01n$ modes) and values of n from 1 to 9. This drawing shows that, for each discrete value of r, minimum W/P is given by points on the "envelope" of the family of curves.

The standard method of obtaining the envelope is to express W as a function of P with n as parameter (r is assumed fixed, for the moment), i.e., $W = F(P, n)$, and then set $\frac{\partial F}{\partial n} = 0$. However, in this case it is easier to express $W = G(P, \varphi)$ and $\varphi = H(n)$, whence

$$\frac{\partial F}{\partial n} = \frac{\partial G}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial n}$$

and the envelope is obtained by setting $\frac{\partial G}{\partial \varphi} = 0$ provided $\frac{\partial \varphi}{\partial n} \neq 0$. We proceed, therefore, as follows.

Assume p is continuous, and solve (4) for p, obtaining:

$$p = \frac{\sin^3 \varphi}{\frac{r}{2\pi P} - \cos^3 \varphi}. \tag{11}$$

Now substitute (11) in (5). This gives W as a function of P and φ:

$$W = \frac{r^3}{4\pi^2}\left[\frac{\sin^2 \varphi}{\cos^2 \varphi\left(\frac{r}{2\pi P} - \cos^3 \varphi\right)}\right]. \tag{12}$$

To solve $\frac{\partial W}{\partial \varphi} = 0$, we differentiate and simplify. This yields

$$5 \cos^3 \varphi - 3 \cos^5 \varphi = \frac{r}{\pi P}. \tag{13}$$

Substituting (13) back into (11) yields

$$p = \frac{2 \sin \varphi}{3 \cos^3 \varphi} \tag{14}$$

The situation so far is that, with P and r assigned, W lies on the envelope and is a minimum when φ satisfies (13); p is then given by (14). Obviously, for (13) to hold, it is necessary that

$$\frac{r}{2\pi P} < 1.$$

To obtain the best value of r, the procedure is to differentiate $W_{\min}$ with respect to r, assuming now that r is continuous. and examine for a mini-

mum. We can, however, first differentiate (12) by setting

$$\frac{dW}{dr} = \frac{\partial W}{\partial r} + \frac{\partial W}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial r}$$

and then substitute from (13). However, when (13) is satisfied, $\frac{\partial W}{\partial \varphi} = 0$. This process yields

$$\frac{dW}{dr} = \frac{r^2}{\pi^2} \frac{(2 - 3\cos^2\varphi)}{9\sin^2\varphi\cos^5\varphi}.$$

This shows $\frac{dW}{dr}$ to be positive, when $\cos^2\varphi < \frac{2}{3}$. Hence $\frac{dW}{dr} = 0$ corresponds to a maximum, rather than a minimum.* If $\cos^2\varphi < \frac{2}{3}$, that is, $\varphi > 35°16'$, then r should be as small as possible. The smallest r is 3.83, for the TE $01n$ modes. For $r = 3.83$, and $\varphi > 35°$, from (13) there is obtained $P > 0.75$.

The analysis thus indicates that, for values of $P = Q\frac{\delta}{\lambda}$ greater than 0.75, the TE $01n$ mode yields the smallest ratio W/P or V/Q.

An interesting and simple relation between fa and R for minimum W/P can easily be derived from the foregoing equations. Substitute (14) back into (6), thereby obtaining

$$k = \frac{4Rr}{3a\cos^3\varphi}. \tag{15}$$

Now use (7) with (15) to eliminate $\cos\varphi$, replace k by $2\pi/\lambda$, and r by 3.83, its numerical value for the TE $01n$ modes. This gives

$$\left(\frac{a}{\lambda}\right)^2 R = 2.23$$

or by substituting $\lambda = \frac{c}{f}$, $c = 3 \times 10^{10}$,

$$(fa)^2 R = 20.1 \times 10^{20}.$$

This useful relation was first discovered by W. A. Edson.

Some further discussion is of interest. It is realized that a number of points have not been taken care of in a manner entirely satisfactory mathematically, but nevertheless important practical results have been obtained. As an example, since p and r can assume only discrete values, there are

* It is for this reason that the determination of the stationary values of $W(r, p, \varphi)$, subject to the constraint $P(r, p, \varphi) =$ constant, by La Grange multipliers fails to yield the desired least value of W/P.

specific situations where some mode other than the $TE\ 01n$ gives a smaller W/P. For example, it may be shown that for P between 0.97 and 1.14 the TE 021 mode yields a smaller W than the TE 013 or TE 014 modes. However, the margin is small, and for larger P, the $TE\ 02n$ modes become progressively poorer.

Limitation on Frequency Range of Tunable Cavity as Set by Ambiguity

In the design of a tunable cylindrical resonant cavity intended for use in the $TE\ 01n$ mode, the requirements on Q may dictate a diameter large enough to sustain $TE\ 02n'$ or $TE\ 03n'$ modes. Also, the range of variation of cavity length may be such that the $TE\ 01(n+1)$ mode is supported. As the cavity is required to tune over a certain range of frequency, the maximum frequency range possible in the $TE\ 01n$ mode without interference from the $TE\ 01(n+1)$† or any $TE\ 02$ or $TE\ 03$ modes is of interest. The interference from the $TE\ 01(n+1)$ limits the useful range of the $TE\ 01n$ by the presence of extraneous responses at more than one dial setting for a given frequency or more than one frequency for a given dial setting. In applications so far made, it has been possible to eliminate extraneous responses from the $TE\ 02$ and $TE\ 03$ modes, but crossings of these modes with the main $TE\ 01n$ mode have not been permitted. No designs have had diameters sufficiently large to support $TE\ 04$ modes.

The desired relations are easily obtained by simple algebraic manipulation of equation (1). For simplicity in presentation of the results, we introduce some symbols applicable to this section only:

$$A = \left[\frac{cr_{\ell,m}}{\pi}\right]^2 \qquad B = \left[\frac{c}{2}\right]^2 = 2.247 \times 10^{20}$$

A_0 = value of A for $TE\ 01n$ modes $= 13.371 \times 10^{20}$

$t = A/A_0$

$x_0 = (a/L)^2$ at low frequency end of useful range of $TE\ 01n$ mode

F = frequency range ratio $= \dfrac{\text{maximum } f}{\text{minimum } f}$.

The values of A and t depend upon the interfering mode under consideration. For the $TE\ 02n$ modes, $A = 44.822 \times 10^{20}$, $t = 3.3522$.

The two typical cases of interest are shown on Fig. 3. For case I, am-

† It is easy to show that the extraneous response from the $TE\ 01(n-1)$ mode is not limiting. The proof depends on the inequality $n^2 > (n+1)(n-1)$.

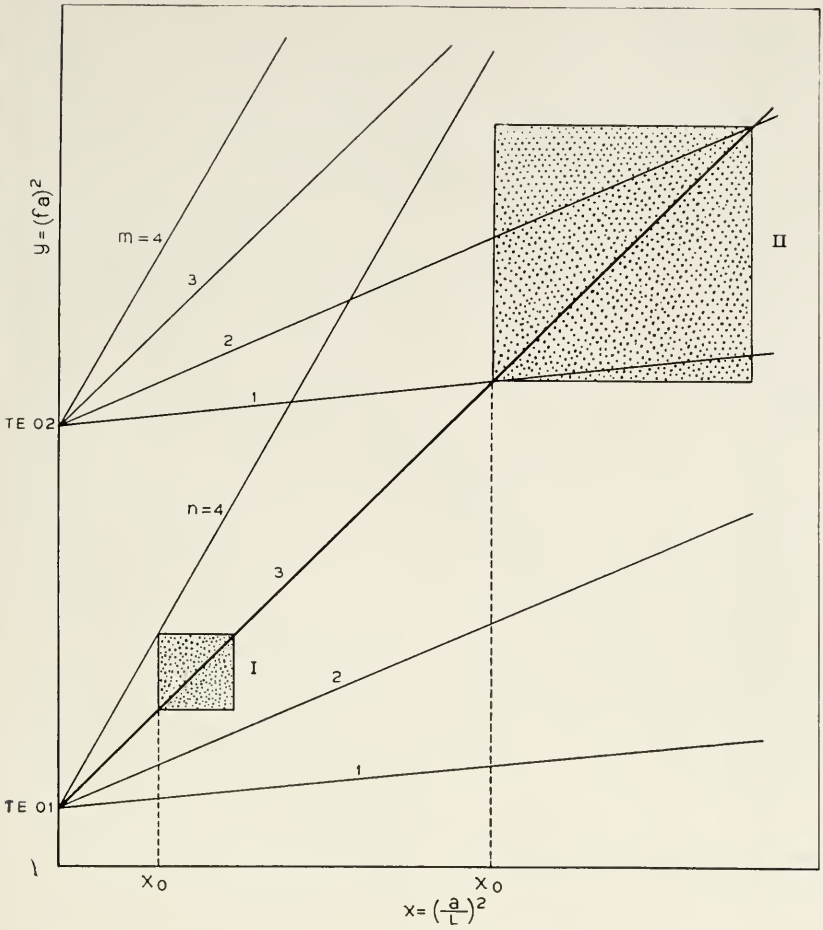


Fig. 3—Mode chart illustrating types of interference with $TE\ 01n$ mode.

biguity from $TE\ 01(n+1)$ mode, it is found that

$$F^2 = \frac{A_0 + B(n+1)^2 x_0}{A_0 + Bn^2 x_0}.$$

Curves of F for this case are shown on Fig. 4.

The maximum value of F is obtained when $x_0 = \infty$ and is

$$F_{\max} = \frac{n+1}{n}.$$

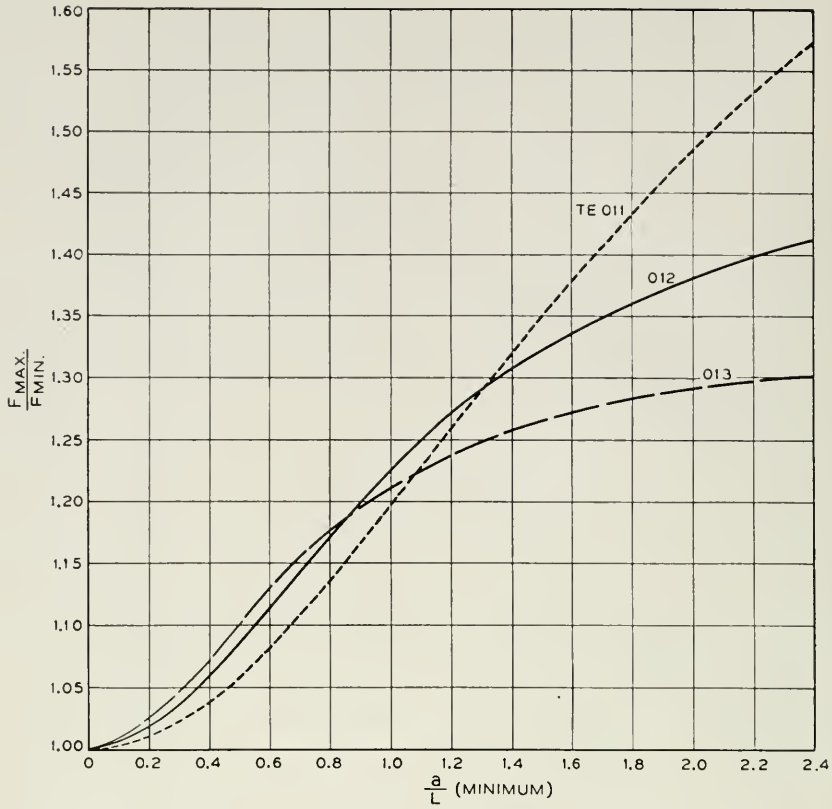


Fig. 4—Curves showing maximum value of frequency ratio without interference from $TE\ 01(n+1)$ mode (case I of Fig. 3).

TABLE I.—*Case II: Maximum Frequency Range Ratio, F, for TE 01n Mode when Limited by Mode Crossings with TE 02m and TE 02(m+1) Modes.*

m	$n=3$		$n=4$		$n=12$	
	F	$(a/L)_{\min}$	F	$(a/L)_{\min}$	F	$(a/L)_{\min}$
1	1.198	1.323	1.086	0.966	1.008	0.313
2			1.242	1.080	1.013	0.316
3					1.019	0.322
4					1.027	0.331
5					1.037	0.343
6					1.051	0.360
7					1.071	0.384
8					1.104	0.418
9					1.168	0.471
10					1.345	0.564

For case II, range limited by mode crossings, it is found that

$$x_0 = \frac{A - A_0}{B(n^2 - n'^2)}$$

$$F^2 = \frac{(n^2 - n'^2)[n^2 t - (n' + 1)^2]}{(n^2 t - n'^2)[n^2 - (n' + 1)^2]}.$$

Some values for this case are given in Table I.

The formulas above are general and may be used for any pair of mode types by using the appropriate values for A and t.

The Elliptic Cylinder

In the design of high Q circular cylinder cavity resonators operating in the $TE\ 01n$ mode, it is desirable to know how much ellipticity is tolerable, so that suitable manufacturing limits may be set. The elliptical wave guide has already been studied, notably by Brillouin[12] and Chu,[13] but the results are not in suitable form or of adequate precision for the present purposes. More recently tables[14] have become available which permit the calculation of some of the properties of the elliptical cylindrical resonator.

The elliptical cavity involves Mathieu functions, which are considerably more complicated than Bessel functions.[16] The tables give the numerical coefficients of series expansions, in terms of sines, cosines, and Bessel functions, of the Mathieu functions up to the fourth order. These tables have been used for the calculation of some quantities of interest in connection with elliptical deformations of a circular cylinder in the $TE\ 01n$ mode.

The Ellipse

All mathematical treatments of the ellipse (including the tables mentioned above) use the eccentricity, e, as the quantity describing the amount of departure from the circular form. The eccentricity is the ratio

$$e = \frac{\text{distance between foci}}{\text{major axis}}.$$

This is not a quantity subject to direct measurement, hence we here introduce and use throughout the ellipticity, E, defined as

$$E = \frac{\text{difference between major and minor diameters}}{\text{major diameter}}.$$

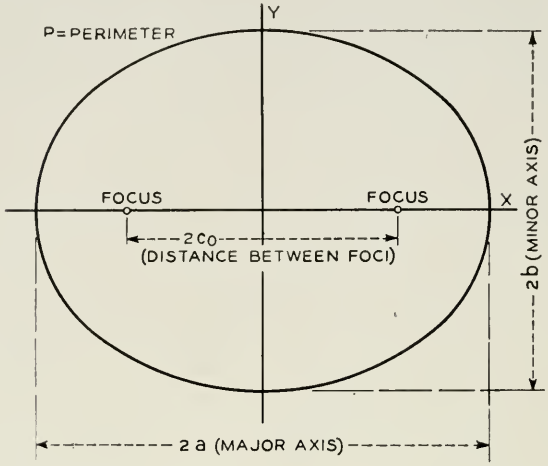
It is clear that the ellipticity is easily obtained directly.

Again, many results are given in terms of the major diameter. Since we are interested in deformations from circular, and in such deformations the

perimeter remains constant, while the major diameter changes, we have expressed our results in terms of an average diameter, defined as

$$D = \frac{\text{perimeter}}{\pi}.$$

Figure 5 shows the ellipse and various relations of interest.



e = ECCENTRICITY = $\frac{c_0}{a}$

E = ELLIPTICITY = $\frac{(a-b)}{a}$

A = AREA = πab

D = "AVERAGE" DIAMETER = $\frac{P}{\pi}$

$b = a\sqrt{1-e^2} = a(1-E)$

$A = \pi a^2\sqrt{1-e^2} = \pi a^2(1-E)$

Fig. 5—The ellipse

Elliptic Coordinates and Functions

The elliptic coordinate system is shown on Fig. 6. Following Stratton,[15] we have used ξ in place of the table's z, since we wish to use z as the coordinate along the longitudinal axis. Stratton also uses $\eta = \cos\varphi$ as the angular coordinate; this is frequently convenient.

Analogous to $\cos \ell\theta$ and $\sin \ell\theta$ in the circular case, there are even and odd* angular functions, denoted by

$$^{e}S_{\ell}(c, \cos\varphi) \text{ and } ^{o}S_{\ell}(c, \cos\varphi)$$

which reduce to $\cos \ell\theta$ and $\sin \ell\theta$ respectively when $c \to 0$. Similarly, there are even and odd* radial functions, denoted by

$$^{e}J_{\ell}(c, \xi) \text{ and } ^{o}J_{\ell}(c, \xi)$$

* For $\ell = 0$, only even functions exist.

which both reduce to $J_\ell(k_1\rho)$ when $c \to 0$. In the above, c is a parameter related to the ellipticity.* The tables do not give values of the functions, but rather give numerical coefficients

$$D_n^\ell \text{ and } F_n^\ell$$

of expansions in series of cosine, sine and Bessel functions, which permit one to calculate the elliptic cylinder functions. The coefficients, of course,

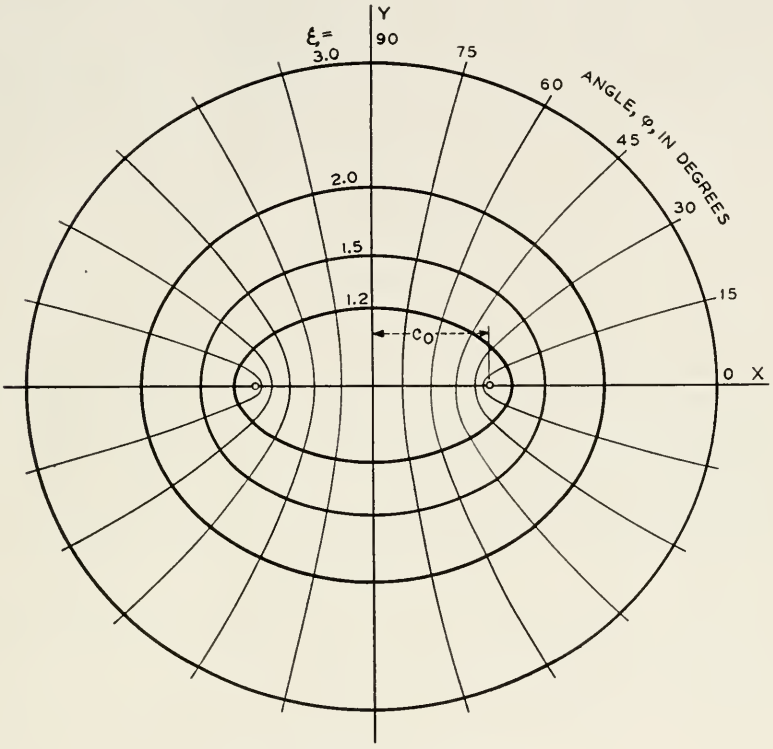


Fig. 6—Elliptic coordinate system

depend on the parameter c; the largest value of c in the tables is 4.5, which corresponds to an ellipticity of 39% in a cylinder operating in the TE $01n$ mode.** For this case, Bessel functions up to $J_{12}(x)$ and $J'_{12}(x)$ are needed for calculating the radial function. It is clear that calculations on elliptic cylinders have not been put on a simple basis.

* Not to be confused with c = velocity of electromagnetic waves; the symbol c is here carried over from the published tables.

** An ellipticity of 39% means that the difference between maximum and minimum diameters is 39% of the maximum diameter. For a given c, the ellipticity depends on the mode.

Field Equations

The equations for the fields are easily obtained from section 6.12 of Stratton's book, and are given in Table II, which is self-explanatory, except for the quantity c, which we now proceed to discuss.

Resonant Frequencies

The elliptic cylinder has the major diameter, $2a$, and the focal distance, $2c_0$. The equation of its surface is then expressed by $\xi = \frac{a}{c_0} = \alpha$. On this surface, E_η must vanish. This requires that ${}^{eo}J'_\ell(c, \alpha) = 0$ for TE modes and that ${}^{eo}J_\ell(c, \alpha) = 0$ for TM modes. The series expansions are in terms of $c\xi$ as variable. Let $c\alpha = r'_{\ell m}$ or $r_{\ell m}$ be the roots of the above equations. Then $\frac{c}{c_0} = \frac{r}{a}$ (dropping the subscripts ℓ, m). Now, in working out the solution of the differential equations, it turned out that $c = c_0 k_1$. Here k_1 is one component of the wave number, k†. Hence $k_1 = \frac{r}{a}$. Furthermore, the eccentricity is $e = \frac{c_0}{a} = \frac{c}{r}$. The indicated procedure is: 1) choose a value of c; 2) find the various values of r for which the radial function or its derivative is zero; 3) then calculate the corresponding eccentricity and resonant frequency. Notice that for a given value of c, the values of r will depend on the mode, and hence so will the eccentricity.

We now wish to express our results in terms of the ellipticity and the average diameter. To convert eccentricity to ellipticity, we use

$$E = 1 - \sqrt{1 - e^2}.$$

The perimeter of the ellipse is given by $P = 4aE(e)$ where $E(e)$ is the complete elliptic integral of the second kind.††
In terms of the average diameter we find

$$k_1 = \frac{2}{D}\left[\frac{2rE(e)}{\pi}\right]$$

or calling the quantity in brackets s, $k_1 = \frac{2s}{D}$. This is now in the same form as k_1 for a circular cylinder of diameter D. The quantity s is the reciprocal of Chu's $\frac{\lambda_c}{S}$.

† It is recalled that

$$k = \frac{2\pi}{\lambda} = \sqrt{k_1^2 + k_3^2}\,; \qquad k_1 = \frac{r}{a}\,; \qquad k_3 = \frac{n\pi}{L}\,.$$

†† This is tabulated as $E(\alpha)$ in Jahnke & Emde, p. 85, with $\alpha = \sin^{-1}e$.

We have calculated and give in Table III values of r, e, E and s for several values of c and for a few modes of special interest. For three cases, ${}^{e}TE\ 01$, ${}^{e}TM\ 11$ and ${}^{o}TM\ 11$, we have determined an empirical formula to fit the calculated values of s. These are also given in Table III.

TABLE II. Elliptic Cylinder Fields

TE Modes

$$E_\xi = -k \sqrt{\frac{\mu}{\epsilon}} \frac{\sqrt{1-\eta^2}}{q} S'_\ell(c,\eta) J_\ell(c,\xi) \sin k_3 z \cos \omega t$$

$$E_\eta = k \sqrt{\frac{\mu}{\epsilon}} \frac{\sqrt{\xi^2-1}}{q} S_\ell(c,\eta) J'_\ell(c,\xi) \sin k_3 z \cos \omega t$$

$$H_\xi = k_3 \frac{\sqrt{\xi^2-1}}{q} S_\ell(c,\eta) J'_\ell(c,\xi) \cos k_3 z \sin \omega t$$

$$H_\eta = k_3 \frac{\sqrt{1-\eta^2}}{q} S'_\ell(c,\eta) J_\ell(c,\xi) \cos k_3 z \sin \omega t$$

$$H_z = k_1^2 S_\ell(c,\eta) J_\ell(c,\xi) \sin k_3 z \sin \omega t$$

TM Modes

$$E_\xi = -k_3 \frac{\sqrt{\xi^2-1}}{q} S_\ell(c,\eta) J'_\ell(c,\xi) \sin k_3 z \cos \omega t$$

$$E_\eta = -k_3 \frac{\sqrt{1-\eta^2}}{q} S'_\ell(c,\eta) J_\ell(c,\xi) \sin k_3 z \cos \omega t$$

$$E_z = k_1^2 S_\ell(c,\eta) J_\ell(c,\xi) \cos k_3 z \cos \omega t$$

$$H_\xi = -k \sqrt{\frac{\epsilon}{\mu}} \frac{\sqrt{1-\eta^2}}{q} S'_\ell(c,\eta) J_\ell(c,\xi) \cos k_3 z \sin \omega t$$

$$H_\eta = k \sqrt{\frac{\epsilon}{\mu}} \frac{\sqrt{\xi^2-1}}{q} S_\ell(c,\eta) J'_\ell(c,\xi) \cos k_3 z \sin \omega t$$

Notes:

Derivatives are with respect to ξ and η.

S_ℓ and J_ℓ carry prefixed superscripts, e or o, since they may be either even or odd.

$$q = c_0 \sqrt{\xi^2 - \eta^2} \qquad c = c_0 k_1$$

$$k_1 = \frac{r_{\ell,m}}{a} \qquad k_3 = \frac{n\pi}{L} \qquad k^2 = k_1^2 + k_3^2$$

$2c_0$ is distance between foci of ellipse.

a is the semi major diameter of the ellipse.

$r_{\ell,m}$ is the value of $c\xi$ that makes

$$J_\ell(c,\xi) = 0 \text{ for } TM \text{ modes}$$

$$J'_\ell(c,\xi) = 0 \text{ for } TE \text{ modes.}$$

TABLE III ROOT VALUES OF RADIAL ELLIPTIC CYLINDER FUNCTIONS

Mode	c	r	e	E	s
^eTE 01	0	3.8317	0	0	3.8317
	0.2	3.8343	0.05216	0.001361	3.8317
	0.4	3.8423	0.10410	0.005434	3.8318
	0.6	3.8558	0.15561	0.012181	3.8324
	0.8	3.8753	0.20643	0.021539	3.8337
	1.0	3.9015	0.25631	0.033406	3.8366
	1.2	3.9349	0.30496	0.047636	3.8417
	1.4	3.9763	0.35209	0.064033	3.8500
	1.6	4.0264	0.39738	0.082346	3.8624
	2.0	4.154	0.4814	0.12351	3.902
	3.0	4.634	0.6474	0.2378	4.101
	4.0	5.29	0.756	0.346	4.42
	4.5	5.66	0.795	0.393	4.62
	$s = 3.8317 + 4.33\,E^2 + 1.9E^3$				
^eTM 11	0	3.8317	0	0	3.8317
	0.2	3.8330	0.05218	0.001362	3.8304
	0.4	3.8370	0.10425	0.005449	3.8265
	0.6	3.8436	0.15610	0.012259	3.8201
	0.8	3.8532	0.20762	0.021791	3.8113
	1.0	3.8658	0.25868	0.034036	3.8003
	1.2	3.8818	0.30913	0.048981	3.7874
	1.4	3.9015	0.35884	0.066599	3.7727
	1.6	3.9253	0.40761	0.086844	3.7568
	4.5	5.13	0.878	0.520	3.91
	$s = 3.8317 - 0.96E + 1.1E^2$				
^oTM 11	0	3.8317	0	0	3.8317
	0.2	3.8356	0.05214	0.001361	3.8330
	0.4	3.8474	0.10397	0.005419	3.8370
	0.6	3.8670	0.15516	0.012111	3.8436
	0.8	3.8944	0.20542	0.021326	3.8530
	1.0	3.9298	0.25446	0.032918	3.8654
	1.2	3.9731	0.30203	0.046701	3.8809
	1.4	4.0243	0.34788	0.062462	3.8997
	$s = 3.8317 + 0.95E + 2.2E^2$				
^eTE 22	0	6.706	0	0	6.706
	0.4	6.712	0.0596	0.00178	6.706
	0.8	6.729	0.1189	0.00709	6.705
	1.2	6.756	0.1776	0.01590	6.702
	1.6	6.788	0.2357	0.02817	6.693
	2.0	6.826	0.2930	0.04389	6.677
^oTE 22	0	6.706	0	0	6.706
	0.4	6.712	0.0596	0.00178	6.706
	0.8	6.730	0.1189	0.00709	6.706
	1.2	6.762	0.1775	0.01587	6.708
	1.6	6.810	0.2350	0.02799	6.715
	2.0	6.877	0.2908	0.04323	6.729

Mode	c	r	e	E	s
^eTE 32	0	8.015	0	0	8.015
	0.4	8.020	0.0499	0.00124	8.015
	0.8	8.035	0.0996	0.00497	8.015
	1.2	8.059	0.1489	0.01115	8.014
	1.6	8.093	0.1977	0.01974	8.013
	2.0	8.135	0.2459	0.03070	8.010
^oTE 32	0	8.015	0	0	8.015
	0.4	8.020	0.0499	0.00124	8.015
	0.8	8.035	0.0996	0.00497	8.015
	1.2	8.060	0.1489	0.01115	8.015
	1.6	8.097	0.1976	0.01972	8.018
	2.0	8.146	0.2455	0.03061	8.022
^eTM 01	0	2.4048	0		
	0.2	2.4090	0.08302		
	0.4	2.4216	0.16518		
	0.6	2.4431	0.24559		
	0.8	2.4739	0.32337		
	1.0	2.5149	0.39762		
^eTE 11	0	1.8412	0		
	0.2	1.8416	0.10860		
	0.4	1.8430	0.21704		
	0.6	1.8452	0.32516		
	0.8	1.8484	0.43280		
	1.0	1.8527	0.53975		

Notes:

Superscripts e and o on mode designation signify even and odd.

c is parameter used in the Tables (Stratton, Morse, Chu, Hutner, "Elliptic Cylinder and Spheroidal Wave Functions")

r is the value of the argument which, for TM modes, makes the radial function zero and, for TE modes, makes its derivative zero.

e is the eccentricity of the ellipse;

$$e = \frac{\text{distance between foci}}{\text{major diameter}}$$

E is the ellipticity of the ellipse;

$$E = \frac{\text{difference between major and minor diam.}}{\text{major diameter}}$$

s is the root value, referred to the "average diameter"; it is related to r by:

$$s = \frac{r}{\pi}\,\frac{\text{perimeter}}{\text{major diameter}}$$

The quantity s is also related to the cutoff wavelength in an elliptical wave guide according to:

$$s = \frac{\text{perimeter of guide}}{\text{cutoff wavelength}}$$

Resonator Q

Although the calculation of the root values is straightforward and not overly laborious, the same cannot be said for the integrations involved in the determination of resonator Q. The procedure is obvious: The field

equations are given; it is only necessary to integrate $H^2 d\tau$ over the volume and $H^2 d\sigma$ over the surface and get Q from

$$Q = \frac{2}{\delta} \frac{\int H^2 d\tau}{\int H^2 d\sigma} \tag{16}$$

with δ = skin depth, a known constant. Unfortunately the integrations cannot at present be expressed in closed form. A numerical solution can be obtained by a combination of integration in series and of numerical integration.

The calculations have been made for the ${}^{e}TE$ 01 mode with $c = 2.0$, for which $r = 4.154$. This value of c corresponds in this case to an ellipticity of about 12%; in a 4″ cylinder this would amount to 1/2″ difference between largest and smallest diameters. Evaluation* of the integrals yields:

$$\int_V H^2 \, d\tau = 12.307 \, k_3^2 L + 12.294 \, k_1^2 L$$

$$\int_S H^2 \, d\sigma = 49.228 \, k_3^2 + 0.1619 \, k_1 k_3^2 L + 6.6847 \, k_1^3 L$$

Substituting $k_1 = \frac{7.804}{D}$ and $k_3 = \frac{\pi n}{L}$, one obtains, finally

$$Q\delta = 0.471 \, D \left(\frac{1 + 0.1622 \, n^2 R^2}{1 + 0.0039 \, n^2 R^2 + 0.1529 \, n^2 R^3} \right).$$

For a circular cylinder,

$$Q_c \delta = 0.5 \, D \left(\frac{1 + 0.1681 \, n^2 R^2}{1 + 0.1681 \, n^2 R^3} \right).$$

Comparison of these two formulas for $Q\delta$ shows that the losses in the end plates ($n^2 R^3$ term) are less with respect to the side wall losses in the elliptical cylinder. The net loss in $Q\delta$, as described by the reduction in the multiplier from 0.5 to 0.471, is thus presumably ascribable to an increase in side wall losses (stored energy assumed held constant). The additional term in $n^2 R^2$ in the denominator is responsible for the difference in the attenuation-frequency behavior of elliptical vs circular wave guide as shown by Chu, Fig. 4. Incidentally, these results agree numerically with those of Chu.

* Numerical integration was by Weddle's rule; intervals of 5° in φ and 0.1 in x were used. The calculations were made by Miss F. C. Larkey.

Corresponding expressions for the resonant wavelength are

$$\lambda = \frac{\pi D}{s\sqrt{1 + \left(\frac{\pi n D}{2sL}\right)^2}} = \frac{0.805\, D}{\sqrt{1 + 0.1622\, n^2 R^2}}$$

$$\lambda_c = \frac{0.820\, D}{\sqrt{1 + 0.1681\, n^2 R^2}}.$$

As an example, take $n = 1$, $R = 1$, then

(Circular)	$Q_c\delta = 0.500\, D$	$\lambda_c = 0.759\, D$
(Elliptical)	$Q\delta = 0.473\, D$	$\lambda = 0.747\, D$
	Ratio = 0.946	Ratio = 0.984.

Conclusions

The mathematics of the elliptic cylinder have not yet been developed to the point where the design of cavities of large ellipticity could be undertaken. On the other hand, sufficient results have been obtained to indicate that the ellipticity in a cavity intended to be circular, resulting from any reasonable manufacturing deviations, would not have a noticeable effect on the resonant frequencies or Q values, at least away from mode crossings.

Full Cylindrical Coaxial Resonator

The full coaxial resonator has been of some interest because of various suggestions for the use of a central rod for moving the tuning piston in a *TE* 01n cavity.

The cylindrical coaxial resonator, with the central conductor extending the full length of the resonator, has modes similar to the cylinder. In fact, the cylinder may be considered as a special case of the coaxial. The indices ℓ, m, n have much the same meaning and the resonant frequencies are determined by the same equation (1). However, now the value of r depends in addition (see Fig. 1) upon η, where

$$\eta = \frac{\text{diameter inner conductor}}{\text{diameter outer conductor}} = \frac{b}{a}.$$

The problem now arises of how best to represent the relations between f, a, b and L. The r's depend on η; so one possibility is to determine their values for a given η and then construct a series of mode charts, one for each value of η.

A more flexible arrangement is to plot the values of r vs η and allow the user to construct graphs suitable for the particular purpose in hand. An equivalent scheme has been used by Borgnis.[18]

It turns out that as $\eta \to 1$, $r(1 - \eta) \to m\pi$, for the *TM* modes and the

TE 0*mn* modes, and $r(1 - \eta) \rightarrow (m - 1)\pi$ for all other *TE* modes. For the former modes, r becomes very large as $\eta \rightarrow 1$, that is, as the inner conductor fills the cavity more and more, the frequency gets higher and higher. For the *TE* $\ell 1n$ modes, however, as the inner conductor grows, the frequency falls to a limiting value. This is discussed in more detail by Borgnis.[18]

Figure 7 shows $r(1 - \eta)$ vs η, for a few of the lower modes; the scale for η between 0.5 and 1.0 is collapsed since this region does not appear to be of great engineering interest. A different procedure is used for the roots of the *TE* $\ell 1n$ modes. Figure 8 is a direct plot of r vs η for a few of the lower modes. In this case, $r \rightarrow \ell$ as $\eta \rightarrow 1$.

Distribution of Normal Modes

The calculation of the distribution of the resonant modes for the coaxial case follows along the lines of that for the cylinder, as given previously. The difference lies in the distribution of the roots r, which now depend upon the parameter η. The determination of this latter distribution offers difficulties. There is some evidence, however, that the normal modes will follow, at least to a first approximation, the same law as the cylinder, viz.:

$$N = 4.4 \frac{V}{\lambda_0^3}$$

with some doubt regarding the value of the coefficient.

$Q \frac{\delta}{\lambda}$ *in Coaxial Resonator*

The integrations needed to obtain this factor are relatively straightforward, but a little complicated. The final results are given in Fig. 1.

The defining equation is (16); the components of H are given in Fig. 1. The integrations can be done with the aid of integrals given by McLachlan[17] and the following indefinite integral:

$$\int \left[\ell \frac{2Z_\ell^2(x)}{x^2} + Z_\ell'^2(x) \right] x\, dx = \frac{x^2}{2} \left[Z_\ell'^2(x) + \frac{2Z_\ell(x)Z_\ell'(x)}{x} + Z_\ell^2(x) \left(1 - \frac{\ell^2}{x^2} \right) \right]$$

which can be verified by differentiation, remembering that $y = Z_\ell(x)$ is a solution of $y'' + \frac{1}{x} y' + \left(1 - \frac{\ell^2}{x^2} \right) y = 0$.

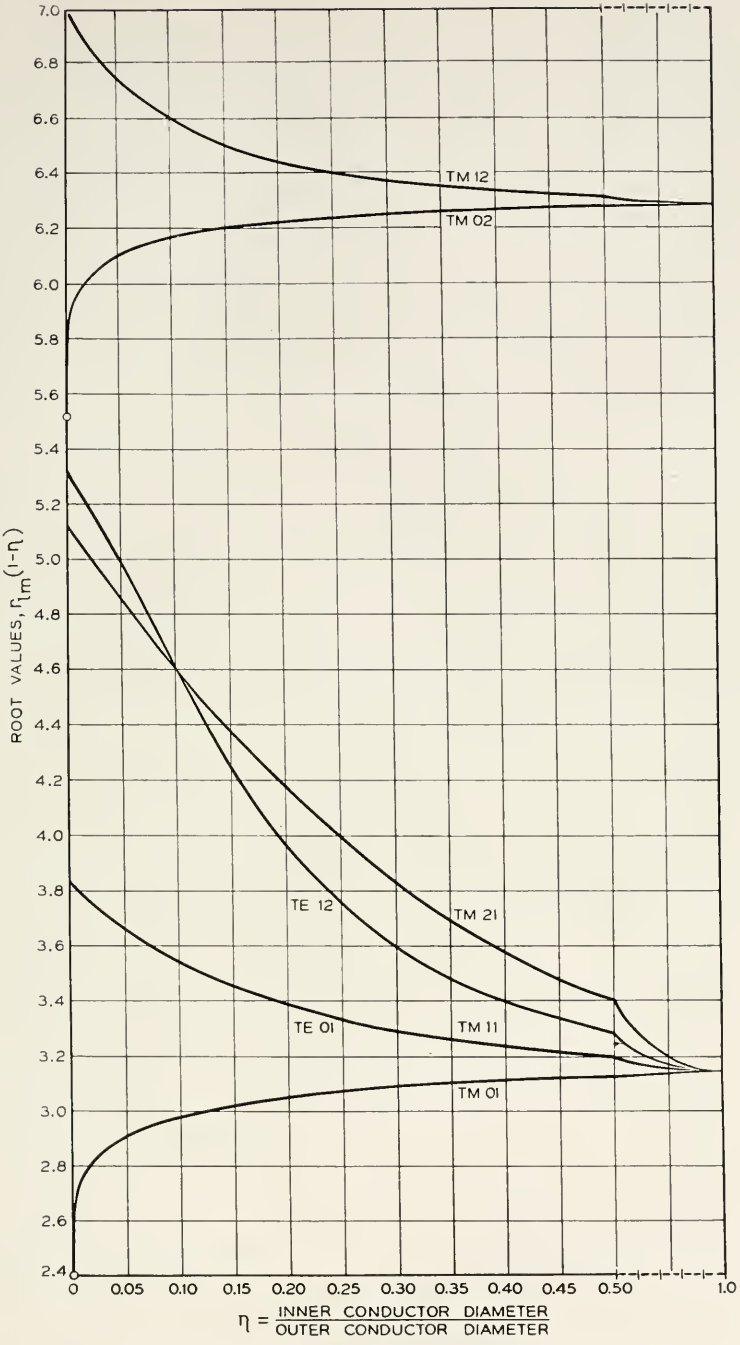


Fig. 7—Full coaxial resonator root values $r_{\ell m}\,(1-\eta)$

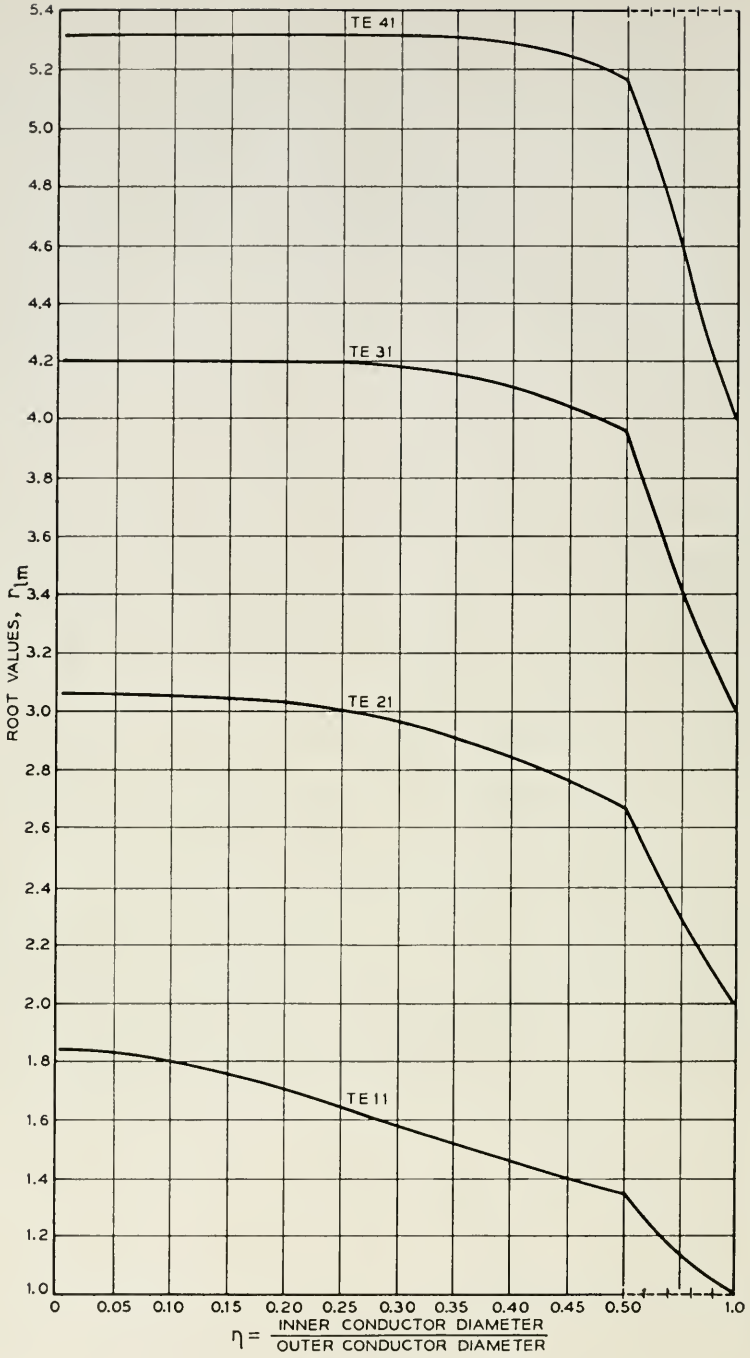


Fig. 8—Full coaxial resonator root values r_{lm}

An investigation needs to be made of the behavior of the formulas as $\eta \to 0$ before any conclusion may be drawn regarding their blending into those for the cylinder. For TE modes with $\ell = 0$, the term involving $\frac{H}{\eta}$ disappears, hence no question arises. Consider then $\ell > 0$, and let $x = \eta r$ for the discussion following. From expansions given in McLachlan, it is easy to show that, for small x

$$Y_\ell(x) \approx -\frac{(\ell-1)!}{\pi}\left(\frac{2}{x}\right)^\ell \qquad Y'_\ell(x) \approx \frac{\ell!}{\pi}\left(\frac{2}{x}\right)^\ell \frac{1}{x}$$

$$J_\ell(x) \approx \frac{x^\ell}{2^\ell \ell!} \qquad J'_\ell(x) \approx \frac{x^{\ell-1}}{2^\ell(\ell-1)!}.$$

Since, from Fig. 1,

$$A = \frac{J'_\ell(r)}{Y'_\ell(r)} = \frac{J'_\ell(\eta r)}{Y'_\ell(\eta r)} = \frac{J'_\ell(x)}{Y'_\ell(x)}$$

it is found, upon substitution of the approximations given above:

$$Z_\ell(x) \approx \frac{2x^\ell}{2^\ell \ell!}.$$

That is, $Z_\ell(x) \sim x^\ell$ and hence $\to 0$ as $x \to 0$. Furthermore $Z_\ell(r)$ remains finite as $\eta \to 0$. Hence $H \sim x^{2\ell}$ and $\frac{H}{\eta} \sim x^{2\ell-1}$. Therefore, for $\ell > 0$, $\frac{H}{\eta} \to 0$ as $\eta \to 0$.

Hence, the expression for $Q\frac{\delta}{\lambda}$ for the coaxial structure reduces to that for the cylinder, for any value of ℓ, in the TE modes.

For the TM modes, and for $\ell > 0$, an entirely similar argument shows that H' remains finite as $\eta \to 0$. Hence, the expression for $Q\frac{\delta}{\lambda}$ for these modes also reduces to that for the cylinder.

For the TM modes, and with $\ell = 0$, we have

$$Z'_0(x) = -J_1(x) + J_0(x)\frac{Y_1(x)}{Y_0(x)}.$$

For $x \to 0$, $J_1(x) \to 0$ and $J_0(x) \to 1$, hence for small x,

$$Z'_0(x) \sim \frac{Y_1(x)}{Y_0(x)}.$$

Now substitute the approximate values of the Y for small x. The result is

$$Z_0'(x) \sim \frac{1}{x \log \frac{x}{2}}.$$

Since $Z_0'(r)$ is finite, it follows that

$$\eta H' \sim \frac{1}{x\left(\log \frac{x}{2}\right)^2}$$

and it is easily shown that $\eta H' \to \infty$ as $\eta \to 0$. On the other hand, $\eta^2 H' \to 0$ as $\eta \to 0$. Hence, $Q\frac{\delta}{\lambda} \to 0$ as $\eta \to 0$. On the other hand, for $\eta = 0$, a

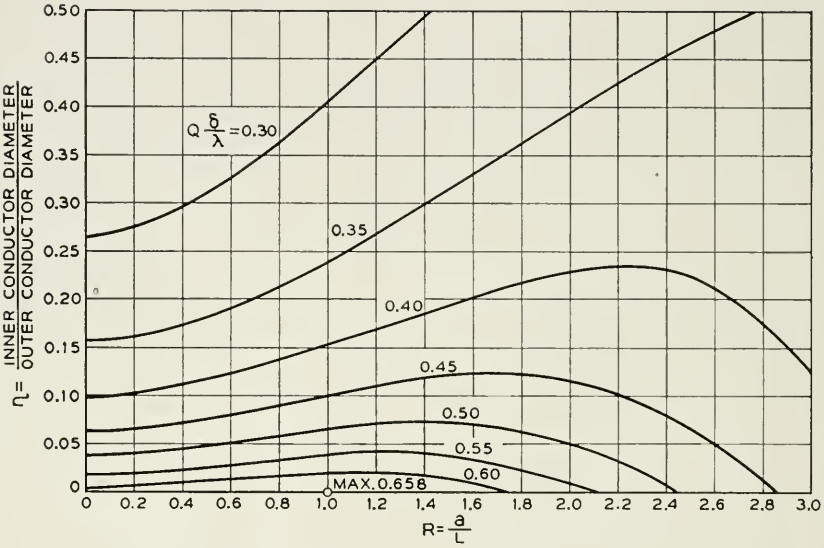


Fig. 9—Coaxial resonator. *TE* 011 mode Contour lines of $Q\frac{\delta}{\lambda}$

perfect cylinder exists whose $Q\frac{\delta}{\lambda}$ is not zero. It is concluded that the expression for $Q\frac{\delta}{\lambda}$ does not apply for small η for the TM modes with $\ell = 0$.

Thus it is seen that the expressions for the factor $(Q\frac{\delta}{\lambda})$ reduce to those given for the cylinder, when $\eta = 0$, except for TM modes with $\ell = 0$. For these latter cases, the factor approaches zero as η approaches zero, because $\eta H'$ increases without limit. This means that an assumption

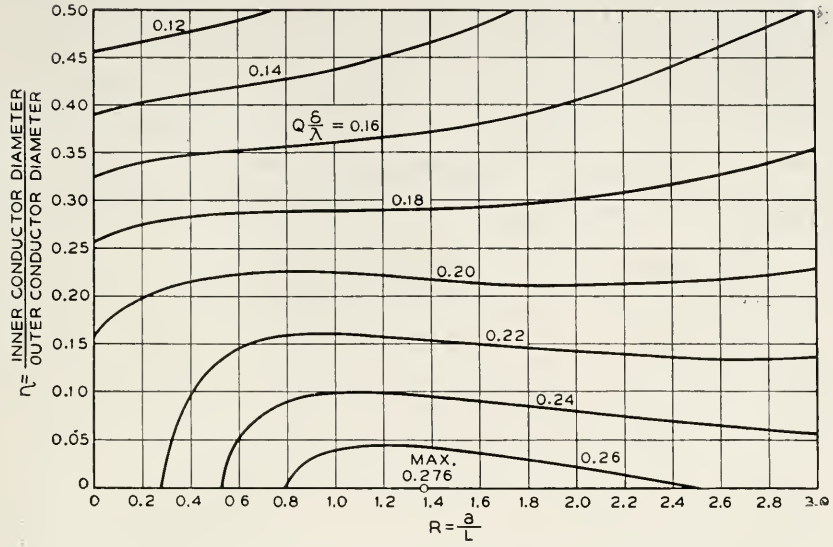


Fig. 10 —Coaxial resonator. *TE* 111 mode Contour lines of $Q\frac{\delta}{\lambda}$

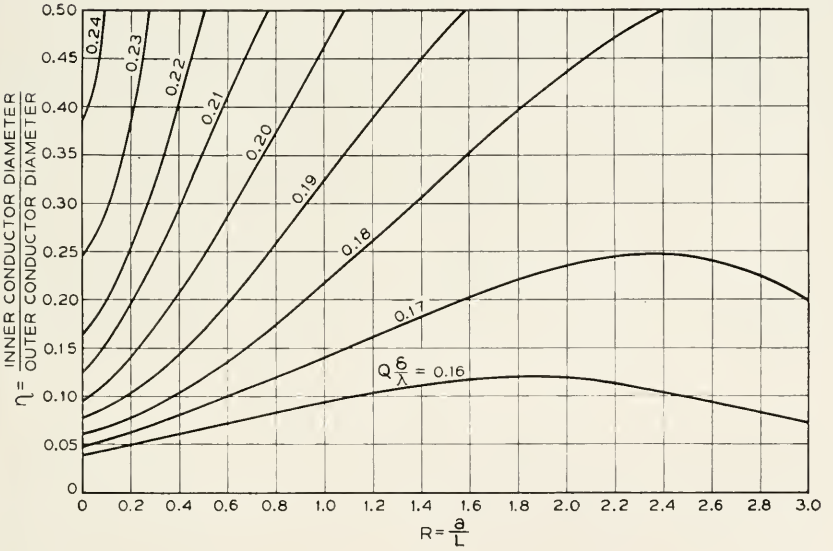


Fig. 11—Coaxial resonator. *TM* 011 mode Contour lines of $Q\frac{\delta}{\lambda}$

which was made in the derivation of the Q values is not valid for small η; that is, the fields for the dissipative case are not the same as those derived on the basis of perfectly conducting walls.

The expressions for the factor are rather complicated, as it depends on several parameters. When a given mode is chosen, the number of parameters reduces to two, η and R. Contour diagrams of $Q\frac{\delta}{\lambda}$ vs η and R are given on Figs. 9, 10, 11 and 12 for the *TE* 011, *TE* 111, *TM* 011 and *TM* 111

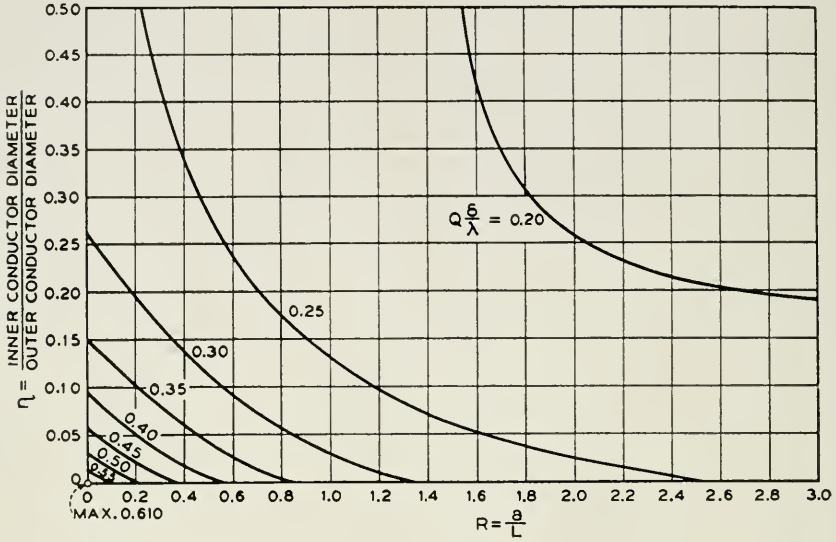


Fig. 12—Coaxial resonator. *TM* 111 mode Contour lines of $Q\frac{\delta}{\lambda}$

modes. As mentioned above, the true behavior of $Q\frac{\delta}{\lambda}$ for the *TM* 011 mode for small η is not given by the above formula, so this contour diagram has been left incomplete.

Fins in a Cavity Resonator

The suppression of extraneous modes is always an important problem in cavity design. Among the many ideas advanced along these lines is the use of structures internal to the cavity.

It is well known that if a thin metallic fin or septum is introduced into a cavity resonator in a manner such that it is everywhere perpendicular to the *E*-lines of one of the normal modes, then the field configuration and

frequency of that particular mode are undisturbed. For example, Fig. 13 shows the E-lines in a TE 11n mode in a circular cylinder. If the upper half of the cylinder wall is replaced by a new surface, shown dotted, the field and frequency in the resulting flattened cylinder will be the same as

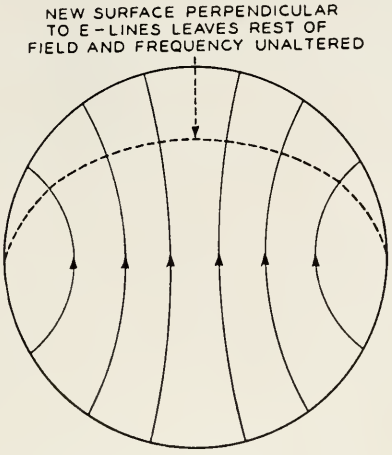


Fig. 13—E Lines in TE 11n mode

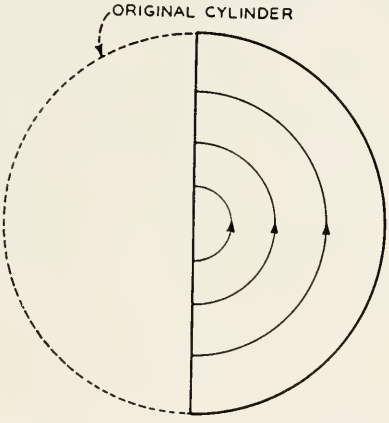


Fig. 14—"TE 01n" mode in half-cylinder

before. Indeed, they will also be the same in the crescent-shaped resonator indicated in the figure.

Except for isolated cases, all the other modes of the original cylinder will be perturbed in frequency since the old fields fail to satisfy the boundary conditions over the new surface. Furthermore, if the original cylinder was

circular, its inherent double degeneracy will be lost and each of the original modes (with minor exceptions) will split into two.

Although the frequency and fields of the undisturbed mode are the same, the Q is not necessarily so. For example, Fig. 14 shows a "*TE* 01*n* mode" in a half cylinder.*

It is easy to calculate $Q\frac{\delta}{\lambda}$ for this case. The result is

$$Q\frac{\delta}{\lambda} = \frac{r}{2\pi} \cdot \frac{(1 + p^2R^2)^{3/2}}{1 + p^2R^3 + K_1 + K_2p^2R^2} \tag{17}$$

in which

$$K_1 = 1.290 \qquad K_2 = 0.653.$$

Here K_1 and K_2 are constants which account for the resistance losses in the flat side. For the full cavity, shown dotted in Fig. 14, eq. (17) holds with $K_1 = K_2 = 0$. If the circular cavity has a partition extending from the center to the rim along the full length, (17) holds with the values of K_1 and K_2 halved. If a fin projects from the rim partway into the interior, still other values of K_1 and K_2 are required. It is a simple matter to compute these for various immersions; Fig. 15 shows curves of K_1 and K_2. The following table gives an idea of the magnitudes involved:

MODE: *TE* 0,1,12 $R = 0.4$

Fin, % a	$Q\frac{\delta}{\lambda}$	Ratio
0%	2.573	1.0
10	2.536	.985
20	2.479	.965
50	2.04	.79
100	1.47	.57

The question now is asked, "Suppose a longitudinal fin were used, small enough to cause only a tolerable reduction in the Q. Would such a fin ameliorate the design difficulties due to extraneous modes?"

Some of the effects seem predictable. All modes with $\ell > 0$ will be split to some extent, into two modes of different frequencies. Consider the *TE* 12*n* mode, for example. There will be one mode, of the same frequency as the original whose orientation must be such that its E-lines are perpendicular to the fin. The Q of this mode would be essentially unchanged. There will be a second mode, oriented generally 90° from the first, whose E-lines will be badly distorted (and the frequency thereby lowered) in the vicinity

* Solutions for a cylinder of this cross-section are known and all the resonant frequencies and Q values could be computed, if they had any application.

of the fin. It would be reasonable to expect the Q of this mode to be appreciably lowered because of the concentrated field there. If two fins at 90° were present, there would be no orientation of the original TE $12n$ mode which would satisfy the boundary conditions. In this case both new modes

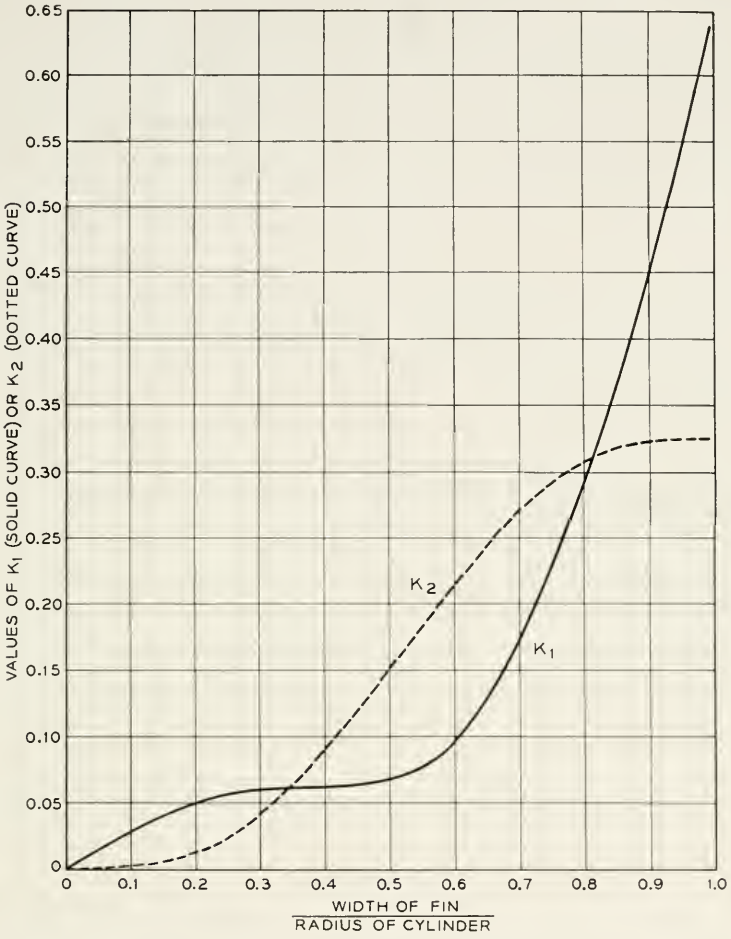


Fig. 15—Constants for calculation of Q of TE $01n$ mode in cylinder with longitudinal fin.

would be perturbed in frequency from the original value. If both fins were identical, the perturbations would be equal and a double degeneracy ensue. Similar effects would happen to the other types of modes.

The major advantage derivable from such effects would appear to be in extraneous transmissions. The fin serves to orient positively the fields in

the cavity, and the input and output coupling locations can then be appropriately chosen. On the basis that internal couplings are responsible for mode crossing difficulties, one might hazard a guess that a real fin would increase such couplings.

Another application of fins might be in a wave guide feed in which it is desired to establish only a $TE\ 0m$ wave. In this case, Q is not so important and larger fins can be used. If these extended virtually to the center and x of them were present (with uniform angular spacing) all types of wave transmission having ℓ less than $x/2$, x even or ℓ less than x, x odd, would be suppressed. This use of fins is an extension of the wires that have been proposed in the past.

Conclusion

It is hoped that the foregoing, which covers some of the theoretical work done by the author during the war, will be of value to other workers in cavity resonators. There is much that needs to be done and hardly time for duplication of effort.

Bibliography

1. E. I. Green, H. J. Fisher, J. G. Ferguson, "Techniques and Facilities for Radar Testing." *B.S.T.J.*, 25, pp. 435–482 (1946).
2. I. G. Wilson, C. W. Schramm, J. P. Kinzer, "High Q Resonant Cavities for Microwave Testing" *B.S.T.J.*, 25, pp. 408–434 (1946).
3. J. R. Carson, S. P. Mead, S. A. Schelkunoff, "Hyper-Frequency Wave Guides—Mathematical Theory," *B.S.T.J.*, 15, pp. 310–333 (1936).
4. G. C. Southworth, "Hyperfrequency Wave Guides—General Considerations and Experimental Results," *B.S.T.J.*, 15, pp. 284–309 (1936).
5. W. W. Hansen "A Type of Electrical Resonator," *Jour. App. Phys.*, 9, pp. 654–663 (1938).—A good general treatment of cavity resonators. Also deals briefly with coupling loops.
6. W. W. Hansen and R. D. Richtmyer, "On Resonators Suitable for Klystron Oscillators," *Jour. App. Phys.*, 10, pp. 189–199 (1939).—Develops mathematical methods for the treatment of certain shapes with axial symmetry, notably the "dimpled sphere," or hour glass.
7. W. L. Barrow and W. W. Mieher, "Natural Oscillations of Electrical Cavity Resonators," *Proc. I.R.E.*, 28, pp. 184–191 (1940). An experimental investigation of the resonant frequencies of cylindrical, coaxial and partial coaxial (hybrid) cavities.
8. R. Sarbacher and W. Edson, "Hyper and Ultrahigh Frequency Engineering," John Wiley and Sons, (1943).
9. R. H. Bolt, "Frequency Distribution of Eigentones in a Three-Dimensional Continuum," *J.A.S.A.*, 10, pp. 228–234 (1939)—Derivation of better approximation formula than the asymptotic one; comparison with calculated exact values.
10. Dah-You Maa, "Distribution of Eigentones in a Rectangular Chamber at Low-Frequency Range," *J.A.S.A.*, 10, pp. 235–238 (1939)—Another method of deriving an approximation formula.
11. I. G. Wilson, C. W. Schramm, J. P. Kinzer, "High Q Resonant Cavities for Microwave Testing," *B.S.T.J.*, 25, page 418, Table III (1946).
12. L. Brillouin, "Theoretical Study of Dielectric Cables," *Elec. Comm.*, 16, pp. 350–372 (1938)—Solution for elliptical wave guides.
13. L. J. Chu, "Electromagnetic Waves in Elliptic Hollow Pipes of Metal," *Jour. App. Phys.*, 9, pp. 583–591 (1938).
14. Stratton, Morse, Chu, Hutner, "Elliptic Cylinder and Spheroidal Wave Functions," M.I.T. (1941).

15. J. A. Stratton, "Electromagnetic Theory," McGraw-Hill, (1941).
16. F. Jahnke and E. Emde, "Tables of Functions," pp. 288–293, Dover Publications (1943).
17. N. W. McLachlan, "Bessel Function for Engineers," Clarendon Press, Oxford (1934).
18. F. Borgnis, "Die konzentrische Leitung als Resonator," *Hochf: tech u. Elek:akus.*, 56, pp. 47–54, (1940).—Resonant modes and Q of the full coaxial resonator. For long abstract, see *Wireless Engineer*, 18, pp. 23–25, (1941).

Additional Bibliography

19. J. J. Thomson, "Notes on Recent Researches in Electricity and Magnetism," Oxford, Clarendon Press, 1893,—§300 gives the resonant frequencies of the *TE* modes in a cylinder with $a/L = 0$; §315–316 consider two concentric spheres; §317–318 treat of the Q of the spherical cavity.
20. Lord Rayleigh, "On the passage of electric waves through tubes or the vibrations of dielectric cylinders" *Phil. Mag.;* 43, pp. 125–132 (1897).—Considers rectangular and circular cross-sections.
21. A. Becker, "Interferenzröhren für elektrische Wellen," *Ann. d. Phys.*, 8, pp. 22–62 (1902)—Abstract in *Sci. Abs.*, 5, No. 1876 (1902)—Experimental work at 5 cm. and 10 cm.
22. R. H. Weber, "Elektromagnetische Schwingungen in Metallröhren," *Ann. d. Phys.*, 8, pp. 721–751 (1902)—Abstract in *Sci. Abs.*, 6A, No. 96 (1903).
23. A Kalähne, "Elektrische Schwingungen in ringförmigen Metallröhren," *Ann. d. Phys.*, 18, pp. 92–127 (1905).—Abstract in *Sci. Abs.*, 8A, No. 2247 (1905).
24. G. Mie, "Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen," *Ann. d. Phys.*, 25, pp. 337–445 (1908)—A part of this article deals with the solution of the equations for the sphere; also shown are the E and H lines for the lowest eight resonant modes.
25. H. W. Droste, "Ultrahochfrequenz-Ubertragung längs zylindrischen Leitern und Nichtleitern," *TFT*, 27, pp. 199–205, 273–279, 310–316, 337–341 (1931)—Abstract in *Wireless Engr.*, 15, p. 617, No. 4209 (1938).
26. W. L. Barrow, "Transmission of Electromagnetic Waves in Hollow Tubes of Metal," *Proc. I.R.E.*, 24, pp. 1298–1328 (1936)—A development of the equations of propagation together with a discussion of terminal connections.
27. S. A. Schelkunoff, "Transmission Theory of Plane Electromagnetic waves," *Proc. I.R.E.*, 25, pp. 1457–1492 (1937)—Treats waves in free space and in cylindrical tubes of arbitrary cross-section; special cases; rectangle, circle, sector of circle and ring.
28. L. J. Chu, "Electromagnetic Waves in Elliptic Hollow Pipes of Metal," *Jour. App. Phys.*, 9, pp. 583–591 (1938)—A study of field configurations, critical frequencies, and attenuations.
29. G. Reber, "Electric Resonance Chambers," *Communications*, Vol. 18, No. 12, pp. 5–8 (1938).
30. F. Borgnis, "Electromagnetische Eigenschwingungen dielektrischer Räume," *Ann. d. Phys.*, 35, pp. 359–384 (1939). Solution of Maxwells equations for rectangular prism, circular cylinder, sphere; also derivations of stored energy and Q values.
31. W. W. Hansen, "On the Resonant Frequency of Closed Concentric Lines," *Jour. App. Phys.*, 10, pp. 38–45 (1939).—Series approximation method for *TM* OOp mode.
32. R. D. Richtmyer, "Dielectric Resonators," *Jour. App. Phys.*, 10, pp. 391–398 (1939).
33. H. R. L. Lamont, "Theory of Resonance in Microwave Transmission Lines with Discontinuous Dielectric," *Phil. Mag.*, 29, pp. 521–540 (1940).—With bibliography covering wave guides, 1937–1939.
34. E. H. Smith, "On the Resonant Frequency of a Type of Klystron Resonator," *Phys. Rev.*, 57, p. 1080 (1940).—Abstract.
35. W. C. Hahn, "A New Method for the Calculation of Cavity Resonators," *Jour. App. Phys.*, 12, pp. 62–68 (1941).—Series approximation method for certain circularly symmetric resonators.
36. E. U. Condon, "Forced Oscillations in Cavity Resonators," *Jour. App. Phys.*, 12 pp. 129–132 (1941).—Formulas for coupling loop and capacity coupling.
37. W. L. Barrow and H. Schaevitz, "Hollow Pipes of Relatively Small Dimensions," *A.I.E.E. Trans.*, 60, pp. 119–122 (1941).—Septate coaxial wave guide and cavity resonator, based on bending a flat rectangular guide into a cylinder.

38. H. König, "The Laws of Similitude of the Electromagnetic Field, and Their Application to Cavity Resonators," *Wireless Engr.*, 19, p. 216–217, No. 1304 (1942). "The law of similitude has strict validity only if a reduction in dimensions by the factor $1/m$ is accompanied by an increase in the conductivity of the walls by the factor m." Original article in *Hochf; tech u. Elek:akus*, 58, pp. 174–180 (1941).
39. S. Ramo, "Electrical Concepts at Extremely High Frequencies," *Electronics*, Vol. 9, Sept. 1942, pp. 34–41, 74–82. A non-mathematical description of the physical phenomena involved in vacuum tubes, cavity resonators, transmission lines and radiators.
40. J. Kemp, "Wave Guides in Electrical Communication," *Jour. I.E.E.*, V. 90, Pt. III, pp. 90–114 (1943).—Contains an extensive listing of U. S. and British patents.
41. H. A. Wheeler, "Formulas for the Skin Effect," *Proc. I.R.E.*, 30, pp. 412–424 (1942)—Includes: a chart giving the skin depth and surface resistivity of several metals over a wide range of frequency; simple formulas for H.F. resistance of wires, transmission lines, coils and for shielding effect of sheet metal.
42. R. C. Colwell and J. K. Stewart, "The Mathematical Theory of Vibrating Membranes and Plates," *J.A.S.A.*, 3, pp. 591–595 (1932)—Chladni figures for a square plate.
43. R. C. Colwell, "Nodal Lines in A Circular Membrane" *J.A.S.A.*, 6, p. 194 (1935)—Abstract.
44. R. C. Colwell, "The Vacuum Tube Oscillator for Membranes and Plates," *J.A.S.A.*, 7, pp. 228–230 (1936)—Photographs of Chladni figures on circular plates.
45. R. C. Colwell, A. W. Friend, J. K. Stewart, "The Vibrations of Symmetrical Plates and Membranes," *J.A.S.A.*, 10, pp. 68–73 (1938).
46. J. K. Stewart and R. C. Colwell, "The Calculation of Chladni Patterns," *J.A.S.A.*, 11, pp. 147–151 (1939).
47. R. C. Colwell, J. K. Stewart, H. D. Arnett, "Symmetrical Sand Figures on Circular Plates," *J.A.S.A.*, 12, pp. 260–265 (1940).
48. V. O. Knudsen, "Resonance in Small Rooms," *J.A.S.A.*, 4, pp. 20–37 (1932)—Experimental check on the values of the eigentones.
49. H. Cremer & L. Cremer, "The Theoretical Derivations of the Laws of Reverberation," *J.A.S.A.*, 9, pp. 356–357 (1938)—Abstract of Akustische Zeits., 2, pp. 225–241, 296–302 (1937)—Eigentones in a rectangular chamber.
50. H. E. Hartig and C. E. Swanson, "Transverse Acoustic Waves in Rigid Tubes," *Phys. Rev.*, 54, pp. 618–626 (1938)—Experimental verification of the presence of acoustic waves in a circular duct, corresponding to the *TE* and *TM* electromagnetic waves; shows an agreement between calculated and experimental values of the resonant frequencies, with errors of the order of $\pm$ 1%.
51. D. Riabouchinsky, *Comptes Rendus*, 207, pp. 695–698 (1938) and 269, pp. 664–666 (1939). Also in Science Abstracts A42, #364 (1939) and A43, #1236 (1940).—Treats of supersonic analogy of the electromagnetic field.
52. F. V. Hunt, "Investigation of Room Acoustics by Steady State Transmission Measurements," *J.A.S.A.*, 10, pp. 216–227 (1939).
53. R. Bolt, "Standing Waves in Small Models," *J.A.S.A.*, 10, p. 258 (1939).
54. L. Brillouin, "Acoustical Wave Propagation in Pipes," *J.A.S.A.*, 11, p. 10 (1939)—Analogy with *TE* waves.
55. P. E. Sabine, "Architectural Acoustics: Its Past and Its Possibilities," *J.A.S.A.*, 11 pp. 21–28, (1939).—Pages 26–28 give an illuminating review of the theoretical work in acoustics.
56. R. H. Bolt, "Normal Modes of Vibration in Room Acoustics: Angular Distribution Theory," *J.A.S.A.*, 11, pp. 74–79 (1939).—Eigentones in rectangular chamber.
57. R. H. Bolt, "Normal Modes of Vibration in Room Acoustics: Experimental Investigations in Non-rectangular Enclosures," *J.A.S.A.*, 11, pp. 184–197 (1939).
58. L. Brillouin, "Le Tuyau Acoustique comme Filtre Passe-Haut," *Rev. D'Acous.*, 8, pp. 1–11 (1939).—A comparison with *TM* waves; some historical notes, tracing the inception of the theory back to 1849.
59. E. Skudrzyk, "The Normal Modes of Vibration of Rooms with Non-Planar Walls," *J.A.S.A.*, 11, pp. 364–365 (1940).—Abstract of Akustische Zeits., 4, p. 172 (1939).—Considers the equivalent of the *TM* OO*p* mode.
60. G. M. Roe, "Frequency Distribution of Normal Modes," *J.A.S.A.*, 13, pp. 1–7 (1941).—A verification of Maa's result for a rectangular room, and an extension to the cylinder, sphere and several derived shapes, which leads to the result that the number of normal modes (acoustic) below a given frequency is the same for all shapes.

61. R. S. Bolt, H. Feshbach, A. M. Clogston, "Perturbation of Sound Waves in Irregular Rooms," J.A.S.A., 14, pp. 65–73 (1942)—Experimental check of eigentones in a trapezoid vs calculated values.

Abstracts of Foreign Language Articles in Wireless Engineer

62. H. Gemperlein, "Measurements on Acoustic Resonators," 16, p. 200, No. 1504 (1939).
63. M. Jouguet, "Natural Electromagnetic Oscillations of a Cavity," 16, p. 511, No. 3873, (1939).
64. M. S. Neiman, "Convex Endovibrators," 17, p. 65, No. 455, (1940).
65. F. Borgnis, "The Fundamental Electric Oscillations of Cylindrical Cavities," 17, p. 112, No. 905, (1940). See also Sci. Abs., B43, No. 343 (1940).
66. H. Buchholz, "Ultra-Short Waves in Concentric Cables, and the "Hollow-Space" Resonators in the Form of a Cylinder with Perforated-Disc Ends," 17, p. 166, No. 1301 (1940).
67. J. Müller, "Investigation of Electromagnetic Hollow Spaces," 17, p. 172, No. 1379 (1940).—Sci. Abs., B43, No. 857 (1940).
68. V. I. Bunimovich, "An Oscillating System with Small Losses," 17, p. 173, No. 1380 (1940).
69. M. S. Neiman, "Convex Endovibrators," 17, p. 218, No. 1743 (1940).
70. M. S. Neiman, "Toroidal Endovibrators," 17, p. 218, No. 1744 (1940).
71. H. Buchholz, "The Movement of Electromagnetic Waves in a Cone-Shaped Horn," 17, p. 370, No. 3009 (1940).—Cavity formed by cone closed by spherical cap.
72. O. Schriever, "Physics and Technique of the Hollow-Space Conductor," 18, p. 18 No. 2, (1941).—Review of history.
73. F. Borgnis, "Electromagnetic Hollow-Space Resonators in Short-Wave Technique," 18, p. 25, No. 61, (1941).
74. T. G. Owe Berg, "Elementary Theory of the Spherical Cavity Resonator," 18, p. 287, No. 1843 (1941).
75. F. Borgnis, "A New Method for measuring the Electric Constants and Loss Factors of Insulating Materials in the Centimetric Wave Band," 18, p. 514, No. 3435 (1941).—An application of the cylindrical cavity resonator.
76. V. I. Bunimovich, "The Use of Rectangular Resonators in Ultra-High-Frequency Technique," 19, p. 28, No. 65 (1942). Use in 17 cm oscillator.
77. V. I. Bunimovich, "A Rectangular Resonator used as a Wavemeter for Decimetric and Centimetric Waves," 19, p. 37, No. 176 (1942).
78. M. Watanabe, "On the Eigenschwingungen of the Electromagnetic Hohlraum," 19, p. 166, No. 927 (1942).
79. F. Borgnis, "The Electrical Fundamental Oscillation of the Cylindrical Two-Layer Cavity," 19, p. 370, No. 2306 (1942). Considers cylindrical resonator with two concentric internal cylinders of different dielectric constant.
80. W. Ludenia, "The Excitation of Cavity Resonators by Saw-Tooth Oscillations," 19, p. 422–423, No. 2641 (1942).
81. Ya. L. Al'pert, "On the Propagation of Electromagnetic Waves in Tubes," 19, p. 520, No. 3181 (1942).—Calculation of losses in a cylindrical wave guide.
82. V. I. Bunimovich, "The Propagation of Electromagnetic Waves along Parallel Conducting Planes," 19, p. 520, No. 3182 (1942).—Equations for Z_0 and attenuation of rectangular wave guide, and resonant frequency and Q of rectangular cavity.
83. C. G. A. von Lindern & G. de Vries, "Resonators for Ultra-High Frequencies," 19, p. 524, No. 3206 (1942).—Discusses transition from solenoid to toroidal coil to "single turn" toroid, i.e., toroidal cavity resonator.

Abstracts in Science Abstracts

84. L. Bouthillon, "Coordination of the Different Types of Oscillations," A39, No. 1773 (1936).—General theory of mechanical, acoustic, optical and electric oscillations.
85. Bürck, Kotowski, and Lichte, "Resonance Effects in Rooms, their Measurement and Stimulation," A39, No. 5226 (1936).
86. G. Jäger, "Resonances of Closed and Open Rooms, Streets and Squares," 40A, No 306 (1937).
87. K. W. Wagner, "Propagation of Sound in Buildings," A40, No. 2199(1937).—Transmission through a small hole in a wall.
88. M. Jouguet, "Natural Electromagnetic Oscillations of a Spherical Cavity," 42A, No. 3822 (1939).
89. H. R. L. Lamont, "Use of the Wave Guide for Measurement of Micro-wave Dielectric Constants," 43A, No. 2684 (1940).

Precision Measurement of Impedance Mismatches in Waveguide

By ALLEN F. POMEROY

A method is described for determining accurately the magnitude of the reflection coefficient caused by an impedance mismatch in waveguide by measuring the ratio between incident and reflected voltages. Reflection coefficients of any value less than 0.05 (0.86 db standing wave ratio) can be measured to an accuracy of ± 2.5%.

LONG waveguide runs installed in microwave systems are usually composed of a number of short sections coupled together. Although the reflection at each coupling may be small, the effect of a large number in tandem may be serious. Therefore, it is desirable to measure accurately the very small reflection coefficients[1] due to the individual couplings.

A commonly adopted method for determining reflection coefficients in phase and magnitude in transmission lines has been to measure the standing wave ratio by means of a traveling detector. Such a system when carefully engineered, calibrated and used is capable of good results, especially for standing waves greater than about 0.3 *db*.

Traveling detectors were in use in the Bell Telephone Laboratories in 1934 to show the reactive nature of an impedance discontinuity in a waveguide. A traveling detector was pictured in a paper[2] in the April 1936 Bell System Technical Journal. Demonstrations and measurements using a traveling detector were included as part of a lecture on waveguides by G. C. Southworth given before the Institute of Radio Engineers in New York on February 1, 1939 and before the American Institute of Electrical Engineers in Philadelphia on March 2, 1939.

Methods for determining the magnitude only of a reflection coefficient by measuring incident and reflected power have been developed by the Bell Telephone Laboratories. A method used during World War II incorporated a directional coupler[3]. The method described in this paper is a refinement of this directional coupler method and is capable of greatly increased accuracy. It uses a hybrid junction[4] to separate the voltage reflected by the mismatch being measured from the voltage incident to the mismatch. Each is measured separately and their ratio is the reflection coefficient.

The problem to be considered is the measurement of the impedance mismatch introduced by a coupling between two pieces of waveguide due to differences in internal dimensions of the two waveguides and to imperfections in the flanges. The basic setup might be considered to be as shown in Fig. 1. The setup comprises a signal oscillator, a hybrid junction, a

calibrated detector and indicator, a termination Z', a piece of waveguide EF (the flange E of which is to be part of the coupling BE to be measured) and a termination Z inserted into the waveguide piece EF so that the reflection coefficient of the coupling BE alone will be measured. In addition a fixed shorting plate should be available for attachment to flange B.

Four cases are considered:

I. Termination Z and Z' perfect, only one coupling on hybrid junction.
II. Termination Z imperfect, termination Z' perfect, only one coupling on hybrid junction.
III. Termination Z perfect, four couplings on hybrid junction.
IV. Termination Z imperfect, four couplings on hybrid junction.

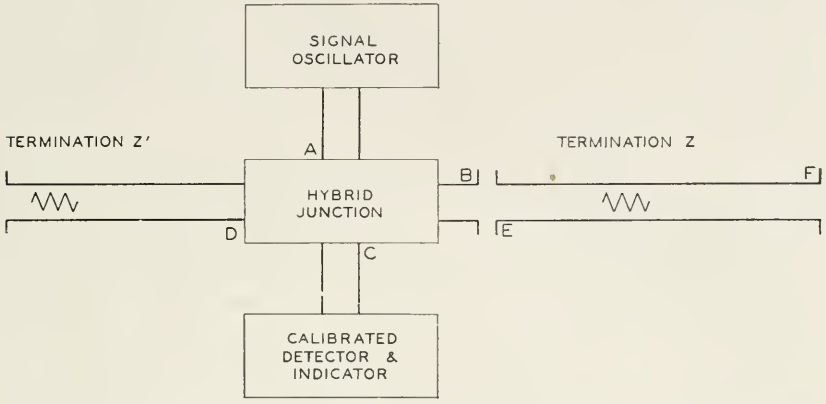


Fig. 1—Block schematic for cases I and II.

It is assumed in all cases that:

1. The hybrid junction has the properties as defined in the discussion of case I.
2. The signal oscillator absorbs all the power reflected through arm A of the hybrid junction.
3. The calibrated detector and indicator absorb all the power transmitted through arm C of the hybrid junction.
4. The oscillator output and frequency are not changed when the hybrid junction arm B is short-circuited.
5. The attenuation of waveguide may be neglected.

I. Termination Z and Z' Perfect, Only One Coupling on Hybrid Junction

In this case the hybrid junction, termination Z' and termination Z, as shown in Fig. 1, are all considered to be perfect. This means for the hybrid junction that its electrical properties are such that the energy from

the oscillator splits equally in paths AD and AB. The half in AD is completely absorbed in the perfect termination Z'. The half in AB is partly reflected from the impedance mismatch due to the waveguide coupling BE and the remainder is absorbed in the perfect termination Z. Again due to the properties of the perfect hybrid junction, the impedance presented by the arm B when arms A and C are perfectly terminated is also perfect, and the reflected energy from waveguide coupling BE splits equally in paths BA and BC. The part in BA is absorbed by the oscillator. The part in BC representing the voltage reflected from the coupling BE is measured by the calibrated detector and indicator. The magnitude of the incident voltage may be measured when the waveguide piece EF is replaced by the fixed shorting plate.

It is convenient to measure voltages applied to the calibrated detector and indicator in terms of attenuator settings in *db* for a reference output indicator reading. Then the ratio expressed in *db* between incident and reflected voltages (hereafter called W) is

$$W_2 \text{ (due to the coupling } BE) = A_1 - A_2 \tag{1}$$

where A_1 is attenuator setting for incident voltage and A_2 is attenuator setting for reflected voltage.

Both reflection coefficient and standing wave ratio may be expressed in terms of W. For if

$$X = \text{voltage due to incident power} \tag{2}$$

and

$$Y = \text{voltage due to reflected power,} \tag{3}$$

$$\text{then reflection coefficient} = \frac{Y}{X} \tag{4}$$

$$\text{and voltage standing wave ratio} = \frac{|X| + |Y|}{|X| - |Y|} \tag{5}$$

$$\text{Since } W(db) = 20 \log_{10} \frac{|X|}{|Y|} \tag{6}$$

$$\text{then in } db\text{, standing wave ratio} = 20 \log_{10} \frac{1 + \text{antilog } \frac{W}{20}}{-1 + \text{antilog } \frac{W}{20}} \tag{7}$$

Standing wave ratio plotted versus W is shown in Fig. 2. Reflection coefficient versus W can be found in any "voltage ratios to *db*" table.

II. Termination Z Imperfect, Termination Z' Perfect, Only One Coupling on Hybrid Junction

In Fig. 1, if the termination Z is not perfect, there will be two reflected voltages from branch B. The vector diagram of the voltage at C might be

represented as in Fig. 3, where vector 0–1 represents the voltage reflected from coupling BE and vector 1–2 represents the voltage reflected from the termination Z. To make measurements, termination Z should be movable

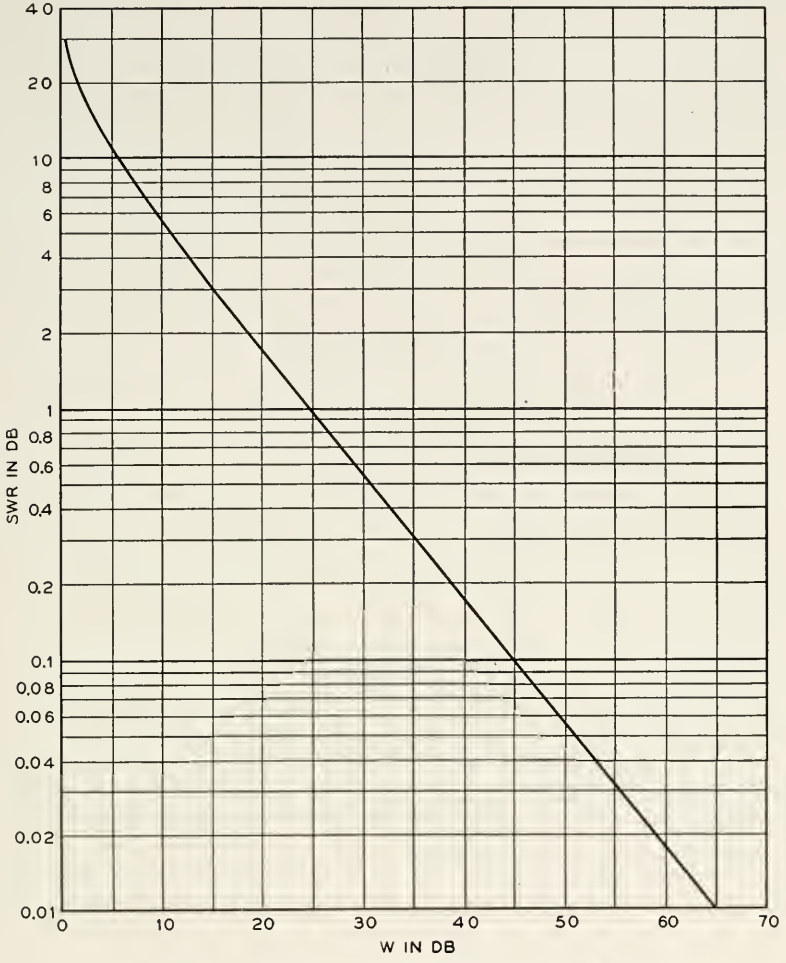


Fig. 2—Standing wave ratio (SWR) versus W.

and the magnitude of its reflection coefficient be the same at a given position of rest for either direction of approach, and be the same for positions of rest over an interval of a half a wavelength in waveguide.

The reflected voltage is measured twice, once for minimum output as the position of the termination Z is adjusted and again for maximum output. Then

$$V_{\min} = V_b - V_z \text{ and } V_{\max} = V_b + V_z \tag{8}$$

where V_b is voltage reflected from coupling BE and V_z is voltage reflected from termination Z.

Equations (8) can be solved for V_b and V_z for

$$V_b = \frac{V_{\max} + V_{\min}}{2} \quad \text{and} \quad V_z = \frac{V_{\max} - V_{\min}}{2} \tag{9}$$

The incident voltage is measured as before. Therefore, using equation (6)

$$W' = 20 \log \frac{|V_a|}{|V_b|} \quad \text{and} \quad W'' = 20 \log \frac{|V_a|}{|V_z|} \tag{10}$$

where W' is due to coupling BE, W'' is due to termination Z and V_a is incident voltage.

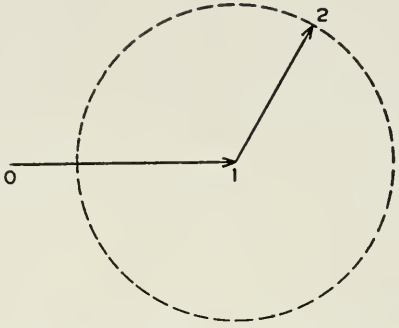


Fig. 3—Vector diagram of voltages reflected from coupling BE and termination Z.

A more practical solution involving only addition, subtraction and the use of the characteristics in Fig. 4 is now presented. The settings of the detector attenuator for incident voltage, minimum output and maximum output might be A_1, A_3 and A_4.

Then $W_3 = A_1 - A_3$ and $W_4 = A_1 - A_4$ (11)

$$\text{But } W_3 = 20 \log \frac{|V_a|}{|V_b| - |V_z|} \quad \text{and} \quad W_4 = 20 \log \frac{|V_a|}{|V_b| + |V_z|} \tag{12}$$

$$\text{and } W_3 - W_4 = 20 \log \frac{|V_b| + |V_z|}{|V_b| - |V_z|} = 20 \log \frac{1 + \text{antilog } \frac{T}{20}}{-1 + \text{antilog } \frac{T}{20}} \tag{13}$$

$$\text{where } 20 \log \frac{|V_b|}{|V_z|} = T = W'' - W' \tag{14}$$

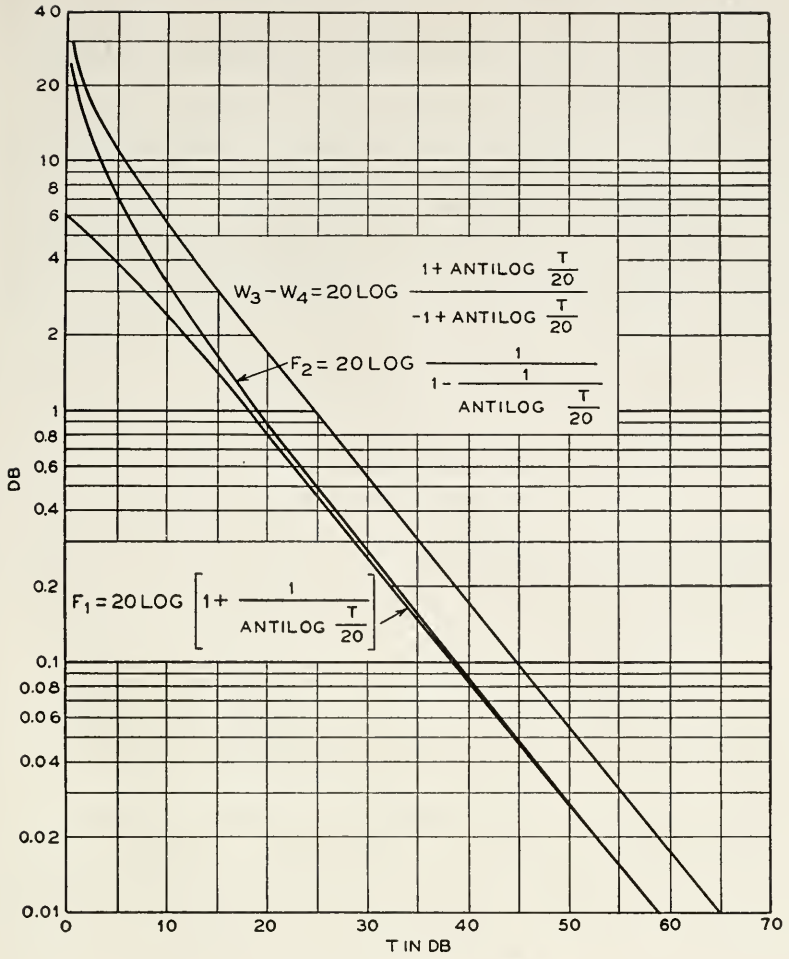


Fig. 4—F_1, F_2 and $W_3 - W_4$

There is an $F_1(T) = 20 \log \left(1 + \dfrac{1}{\text{antilog}\dfrac{T}{20}}\right)$

and an $F_2(T) = 20 \log \dfrac{1}{1 - \dfrac{1}{\text{antilog}\dfrac{T}{20}}}$ (15)

such that $W' = W_4 + F_1 = W_3 - F_2$

$W'' = T + W_4 + F_1 = T + W_3 - F_2$ (16)

and $F_1 + F_2 = W_3 - W_4$

Figure 4 shows F_1, F_2 and their sum $W_3 - W_4$ plotted versus T. It may be noted that $W_3 - W_4$ versus T has the same values as SWR versus W in Fig. 2.

Using equations (16) and Fig. 4, W' and W'' may be evaluated for the particular values of W_3 and W_4 in equation (11). In the evaluation, if there is uncertainty as to which reflection coefficient belongs to the waveguide coupling BE and which belongs to the termination Z, a termination with a different magnitude of reflection coefficient should be used and the technique repeated. The reflection coefficient which is the same in the two cases is of course that due to the waveguide coupling BE.

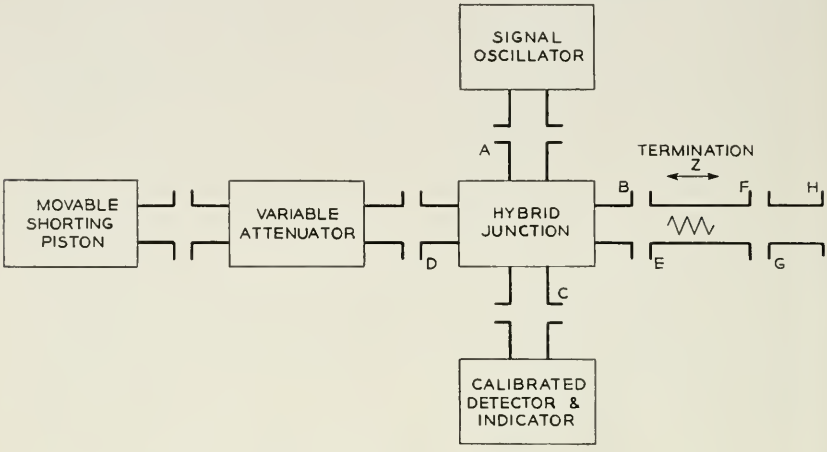


Fig. 5—Block schematic for cases III and IV.

It is assumed in the above solution that multiple reflections between the two impedance mismatches are inconsequential. Appendix A outlines a procedure for evaluating the maximum probable error due to multiple reflections.

III. Termination Z Perfect, Four Couplings on Hybrid Junction

In this case the setup might be as shown in Fig. 5. This setup differs from that shown in Fig. 1 in that the hybrid junction has four couplings shown, termination Z' has been replaced by a variable attenuator and a movable shorting piston, and the waveguide coupling FG is to be measured instead of coupling BE. The hybrid junction and the termination Z are assumed to be perfect as defined for case I.

Since it is the object of the measuring method to measure impedance mismatches in branch B, it is desirable to make the voltage at C depend only on power reflected from branch B. This is accomplished by adjusting

branch D so that the voltages due to the flanges of the hybrid junction are cancelled.

The vector diagram of the voltage at C might be represented as in Fig. 6. Vector 0–1 represents the voltage at C when input is applied to A, due to the impedance mismatch at the coupling BE. Vector 1–2 represents that due to the mismatch at coupling D. Vector 2–3 represents that due to the mismatch at the variable attenuator, (which will usually change in magnitude and probably in phase for different settings). Vector 3–0 represents the voltage at C due to the cancelling voltage from the branch D. Its phase can be varied by changing the position of the movable shorting piston. Its magnitude can be varied by changing the setting of the variable attenuator. When the adjustment is accomplished effectively no power reaches the detector. It is necessary that the reflection coefficients of

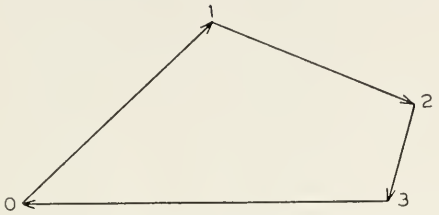


Fig. 6—Vector diagram of voltages at terminal C.

couplings A, B, and C be small so that multiple reflections caused by them will not affect the accuracy of measurement.

The reflected power from coupling FG may be measured when waveguide GH is connected to waveguide EF as shown in Fig. 5 and termination Z is located within waveguide GH. The detector attenuator setting might be A_5. The incident power may be measured as before when termination Z is withdrawn from the waveguide EF and the piece of waveguide GH is replaced by a fixed shorting plate.

$$W_5 \text{ (due to reflection coefficient of the coupling } FG) = A_1 - A_5 \quad (17)$$

IV. Termination Z Imperfect, Four Couplings on Hybrid Junction

In Fig. 5 if the movable termination Z is not perfect, there will be two reflected voltages in branch B when the adjustment is being made. The vector diagram of the voltage at C might be as in Fig. 7. This is the same as Fig. 6 except that a new vector 0–5 represents the voltage due to the mismatch of the movable termination Z. The adjustment is accomplished the same as in the last section except that the criterion is to have no change in detector output as the movable termination Z is moved axially over a

range of a half a wavelength in waveguide. As for the last case it is necessary that the reflection coefficients of the couplings A, B and C be small if good accuracy is desired.

When measuring the coupling FG the procedure and evaluation are the same as for case II.

Part of a laboratory setup as used at about 4 kilomegacycles is shown in Fig. 8. It includes a hybrid junction, a variable attenuator, a movable shorting piston, a straight section of waveguide and a movable termination which consists of a cylinder of phenol resin and carbon with a tapered section at one end. It is mounted in a phenolic block so that it may be moved axially in the wave guide.

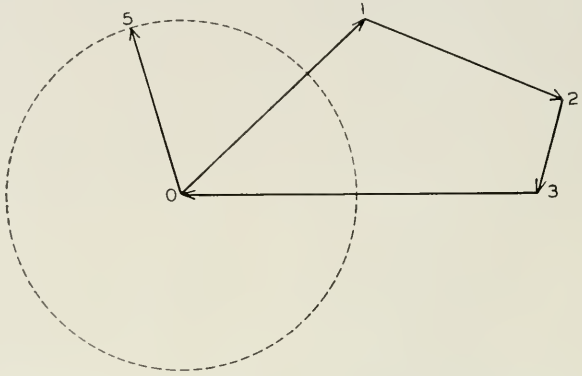


Fig. 7—Part of a laboratory setup as used at 4 kilomegacycles.

In cases III and IV if the hybrid junction has "poor balance" so that voltage appears at C when input is applied to arm A even though B and D are perfectly terminated, the adjusting procedure will cancel this voltage as well. Measuring accuracy will not be impaired provided the other assumptions are fulfilled.

Measuring W—A Fitting Which Does Not Admit of Measuring Each End Separately

A piece with a configuration unsuited to the preceding technique may be measured by connecting it between two straight pieces of waveguide such as between flanges F and G in Fig. 5. The W due to the vector sum of the reflection coefficients of the coupling at one end, any irregularities and the coupling at the other end, is measured. Due to the distance between the mismatches, the vector sum will vary over the frequency band of interest.

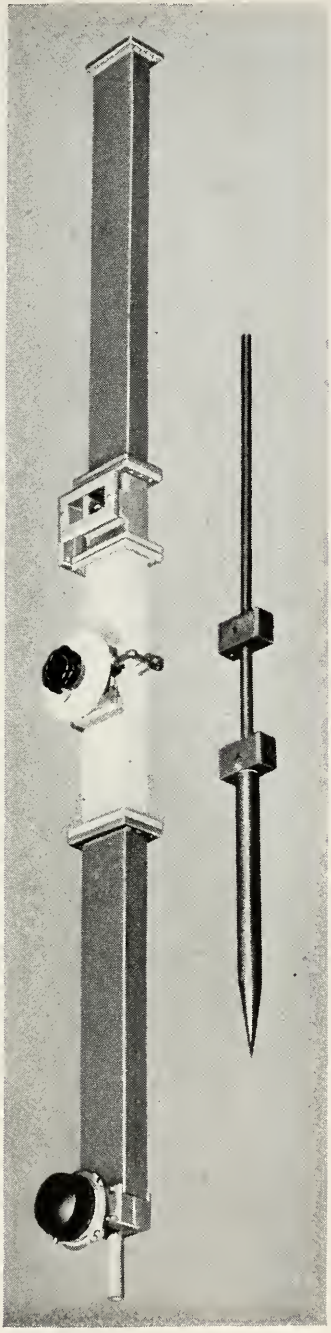


Fig. 8—Vector diagram of voltages at terminal C when termination Z is not perfect.

Accuracy

There are three important sources of error. The first is lack of proper adjustment. The second is that due to the detector attenuator calibration. The third is that due to multiple reflections.

Experience and care can almost eliminate the first source. The second source may have a magnitude of twice the detector attenuator calibration error. In equations (1) and (17) this is readily apparent. The evaluation of W using equations (16) introduces negligibly more error provided $W_3 - W_4$ is made large by proper choice of the magnitude of the reflection coefficient of the termination Z. The possible errors due to multiple reflections between the waveguide impedance discontinuity being measured and an imperfect termination are discussed in Appendix A. If the impedance presented by the arm B of the hybrid junction is not perfect, energy reflected from the hybrid junction will be partly absorbed in the termination and cause an error in the measurement. If the magnitude of this reflection coefficient is known, the maximum error may be computed.

If a detector attenuator calibration error of ± 0.1 *db* is assumed to be the only contributing error, it is possible to measure the W due to an impedance mismatch to an accuracy of ± 0.2 *db* provided the W is greater than 26 *db*. These numbers correspond to measuring a standing wave ratio of any value less than 0.86 *db* to an accuracy of ± 0.02 *db* or reflection coefficients of any value less than 0.05 to an accuracy of $\pm 2.5\%$.

APPENDIX A

Maximum Probable Error Due to Magnitude of Reflection Coefficient Being Measured When Measuring a Waveguide Coupling

The purpose of this appendix is to derive equations so that the maximum probable error due to multiple reflections may be calculated. The assumptions may not be rigorous but the mathematical treatment appears to represent a reasonable approximation. It is assumed that there is no dissipation in waveguide *EF*, waveguide *GH* and in coupling *FG*.

The electrical relations of the coupling *FG* and the movable termination Z might be represented as in Fig. 9, where K_a = characteristic impedance of waveguide *EF* and K_b = characteristic impedance of waveguide *GH*. The first few multiple reflections from the two discontinuities, coupling *FG* and termination Z, can be illustrated as in Fig. 10.

Evaluation of the magnitudes of the reflections can be accomplished as outlined in paragraph 7.13, page 210 in the book "Electromagnetic Waves"* by S. A. Schelkunoff.

* Published by D. Van Nostrand, Inc., New York City, 1943.

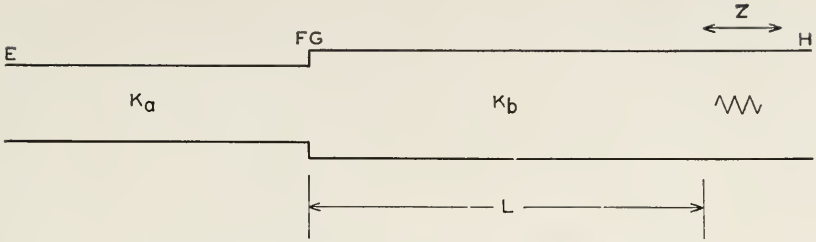


Fig. 9—Relation between coupling *FG* and termination *Z*.

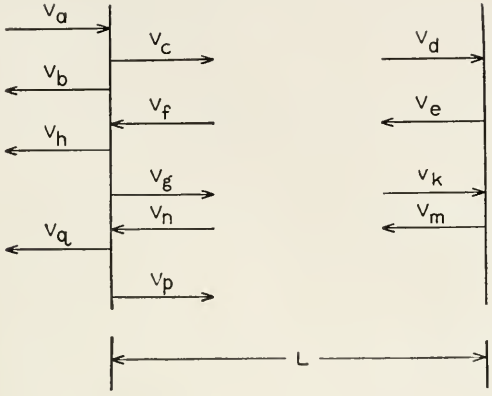


Fig. 10—Multiple reflections from two planes of discontinuity.

$$V_a = \text{Incident voltage} \tag{18}$$

$$V_b = rV_a \tag{19}$$

$$\text{where } r = \frac{K_b - K_a}{K_b + K_a} \tag{20}$$

$$V_c = V_a + V_b = V_a(1 + r) \tag{21}$$

$$V_d = e^{-i\beta L} V_c = e^{-i\beta L} V_a(1 + r) \tag{22}$$

$$V_e = zV_d = ze^{-i\beta L} V_a(1 + r) \tag{23}$$

$$\text{where } z = \frac{Z - Z_b}{Z + Z_b} \tag{24}$$

$$V_f = e^{-i\beta L} V_e = ze^{-i2\beta L} V_a(1 + r) \tag{25}$$

$$V_g = -rV_f = ze^{-i2\beta L} V_a(1 + r)(-r) \tag{26}$$

where $-r = \dfrac{K_a - K_b}{K_a + K_b}$ (27)

$$V_h = V_f + V_g = ze^{-i2\beta L} V_a(1 + r)(1 - r) \quad (28)$$

$$V_k = e^{-i\beta L} V_g = ze^{-i3\beta L} V_a(1 + r)(-r) \quad (29)$$

$$V_m = zV_k = z^2 e^{-i3\beta L} V_a(1 + r)(-r) \quad (30)$$

$$V_n = e^{-i\beta L} V_m = z^2 e^{-i4\beta L} V_a(1 + r)(-r) \quad (31)$$

$$V_p = -rV_n = z^2 e^{-i4\beta L} V_a(1 + r)(-r)^2 \quad (32)$$

$$V_q = V_n + V_p = z^2 e^{-i4\beta L} V_a(1 - r^2)(-r) \quad (33)$$

For purposes of analysis it is now assumed that further multiple reflections are negligible.

Fig. 11—Vector voltage diagram for maximum vector sum.

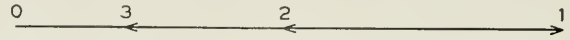


Fig. 12—Vector voltage diagram for minimum vector sum.

Equations (19), (28) and (33) are the reflected voltages that combine vectorially to be measured. If $\beta L = 0, \pi, 2\pi, \cdots n\pi$ then the vector voltage diagram might appear as in Fig. 11. If $BL = \dfrac{\pi}{2}, \dfrac{3\pi}{2}, \dfrac{5\pi}{2}, \cdots \dfrac{(2n-1)\pi}{2}$ then the vector voltage diagram might appear as in Fig. 12.

The following example illustrates the calculations involved in computing the errors due to the magnitude of the reflection coefficient being measured. The assumptions are such that an appreciable error is computed. If one assumes $r = 0.316$ and $z = 0.282$, then from equation (6) $W_r = 10$ *db* and $W_z = 11$ *db*. In Figs. 11 and 12,

$$\text{vector } 0\text{–}1 = r, \text{ vector } 1\text{–}2 = z(1 - r^2), \text{ vector } 2\text{–}3 = rz^2(1 - r^2) \quad (34)$$

then

$$W_{0-1} = 10\ db, W_{1-2} = 11.00 + 0.92 = 11.92\ db,$$
$$\text{and } W_{2-3} = 10.00 + 22.00 + 0.92 = 32.92\ db \quad (35)$$

In order to evaluate vector 0–2 in Fig. 11 (the vector sum of vectors 0–1 and 1–2), one calculates their difference T.

$$T = 11.92 - 10.00 = 1.92\ db \quad (36)$$

$$\text{For } T = 1.92\ db, F_1 = 5.10\ db \quad (37)$$

$$\text{therefore } W_{0-2} = 10.00 - 5.10 = 4.90\ db \quad (38)$$

In order to evaluate vector 0–3 in Fig. 11 (the vector difference of vectors 0–2 and 2–3) one calculates their difference T.

$$T = 32.92 - 4.90 = 28.02\ db \tag{39}$$

$$\text{For } T = 28.02\ db,\ F_2 = 0.36\ db \tag{40}$$

$$\text{therefore } W_{0-3} = 4.90 \pm 0.36 = 5.26\ db = W_4 \tag{41}$$

In order to evaluate vector 0–2 in Fig. 12 (the vector difference between vectors 0–1 and 1–2), one uses T from equation (36).

$$\text{For } T = 1.92\ db,\ F_2 = 14.10\ db \tag{42}$$

$$\text{therefore } W_{0-2} = 10.00 + 14.10 = 24.10\ db \tag{43}$$

In order to evaluate vector 0–3 in Fig. 12 (the vector difference between vectors 0–2 and 2–3), one calculates their difference T.

$$T = 32.92 - 24.10 = 8.82\ db \tag{44}$$

$$\text{For } T = 8.82\ db,\ F_2 = 3.93\ db \tag{45}$$

$$\text{therefore } W_{0-3} = 24.10 + 3.93 = 28.03\ db = W_3 \tag{46}$$

Using equation (16)

$W_3 - W_4 = 22.77\ db$, $T = 1.24\ db$, $F_1 = 5.40$ and therefore $W = 9.66\ db$.

Since we started by assuming $W_r = 10\ db$, the error amounts to 0.34 db.

References

1. Page 120, "Transmission Networks and Wave Filters," T. E. Shea. Published by D. Van Nostrand, Inc., New York City, 1929.
2. "Hyper-frequency Waveguides—General Considerations and Experimental Results," G. C. Southworth, *Bell System Technical Journal*, April, 1936.
3. "Directional Couplers." W. W. Mumford, *Proceedings of the Institute of Radio Engineers*, February 1947.
4. "Hybrid Circuits for Microwaves," W. A. Tyrrell. A paper accepted for publication in the *Proceedings of the Institute of Radio Engineers*.
5. "Note on a Reflection-Coefficient Meter," Nathaniel I. Korman, *Proceedings of the Institute of Radio Engineers and Waves and Electrons*, September 1946.
6. "Probe Error in Standing-Wave Detectors," William Altar, F. B. Marshall and L. P. Hunter, *Proceedings of the Institute of Radio Engineers and Waves and Electrons*, January 1946.
7. Pages 20 to 24, "Practical Analysis of UHF Transmission Lines—Resonant Sections—Resonant Cavities—Waveguides," J. R. Meagher and H. J. Markley Pamphlet published by R. C. A. Service Company, Inc., in 1943.
8. "Microwave Measurements and Test Equipments," F. J. Gaffney, *Proceedings of the Institute of Radio Engineers and Waves and Electrons*, October 1946.

Reflex Oscillators

By J. R. PIERCE and W. G. SHEPHERD

Table of Contents

I. Introduction 463
II. Electronic Admittance—Simple Theory 467
III. Power Production for Drift Angle of $(n + \frac{3}{4})$ Cycles 470
IV. Effect of Approximations 479
V. Special Drift Fields 480
VI. Electronic Gap Loading 482
VII. Electronic Tuning—Arbitrary Drift Angle 484
VIII. Hysteresis 493
IX. Effect of Load 512
A. Fixed Loads 513
B. Frequency Sensitive Loads—Long Line Effect 523
C. Effect of Short Mismatched Lines on Electronic Tuning 531
X. Variation of Power and Electronic Tuning with Frequency 537
XI. Noise Sidebands 542
XII. Build-up of Oscillation 545
XIII. Reflex Oscillator Development at the Bell Telephone Laboratories 550
A. Discussion of the Beating Oscillator Problem 550
B. A Reflex Oscillator with an External Resonator—The 707 553
C. A Reflex Oscillator with an Integral Cavity—The 723 558
D. A Reflex Oscillator Designed to Eliminate Hysteresis—The 2K29 563
E. Broad Band Oscillators—The 2K25 570
F. Thermally Tuned Reflex Oscillators—The 2K45 577
G. An Oscillator with Wave-Guide Output—The 2K50 597
H. A Millimeter—Range Oscillator—The 1464 603
I. Oscillators for Pulsed Applications—The 2K23 and 2K54 607
J. Scope of Development at the Bell Telephone Laboratories 620

Appendices

I. Resonators 622
II. Modulation Coefficient 629
III. Approximate Treatment of Bunching 639
IV. Drift Angle as a Function of Frequency and Voltage 643
V. Electronic Admittance—Non-simple Theory 644
VI. General Potential Variation in the Drift Space 656
VII. Ideal Drift Field 660
VIII. Electronic Gap Loading 663
IX. Losses in Grids 673
X. Starting of Pulsed Reflex Oscillators 674
XI. Thermal Tuning 677

Symbols

A	A measure of frequency deviation (9.20).
B	Bandwidth (Appendix 10 only, (j-3)).
B	Susceptance
B_1	Reduced susceptance (9.7).
B_e	Electronic susceptance.
C	Capacitance
C	Heat capacity (k-1).
D	Reduced gap spacing (10.3).

E_0 Retarding field in drift space.
F Drift effectiveness factor (5.4).
G Conductance
G_1, G_2 Reduced conductances (9.6), (9.12).
G_E Gap conductance of loaded resonator.
G_e Electronic conductance.
G_L Conductance at gap due to load.
G_R Conductance at gap due to resonator loss.
H Efficiency parameter (3.7).
H_m Maximum value of H for a given resonator loss.
I Radio-frequency current.
I_2 Current induced in circuit by convection current returning across gap.
I_0 D-C beam current.
K Resonator loss parameter (3.9).
K Radiation loss in watts/(degree Kelvin)^4 (k-2).
L Inductance.
M Characteristic admittance (a-8).
M_L Characteristic admittance of line.
M_M Short line admittance parameter (9.38).
N Drift time in cycles.
N Length of line in wavelengths (Section IX only).
N Transformer voltage ratio.
P Power.
Q Equation (a-10).
Q_E External Q (a-11).
Q_0 Unloaded Q (a-12).
R Surface resistance (a-2).
S Scaling factor (9.17).
T Temperature.
V Radio-frequency voltage.
V Potential in drift space (Appendix VI only).
V_0 D-C beam voltage at gap.
V_R The repeller is at a potential $(-V_R)$ with respect to the cathode.
W Work, energy (Appendix I).
W Reduced radian frequency (10.5).
X Bunching parameter (2.9).
Y Admittance.
Y_C Circuit admittance.
Y_e Electronic admittance.
Y_L Load admittance.
Y_R Resonator admittance.
Z Impedance.
Z_L Load impedance.
a Distance between grid wire centers.
d Separation between grid planes or tubes forming gap.
e Electronic charge (1.59×10^{-19} Coulombs).
f Frequency.
f Factor relating to effective grid voltage (b-37).
i Radio-frequency convection current.
i_2 Radio-frequency convection current returning across gap (c-4).
$(i_2)_f$ Fundamental component of i_2.
j $\sqrt{-1}$
k Boltzman's constant (1.37×10^{23} joules/degree K).
k Conduction loss watts/degree C (k-14).
l Length.
m Mutual inductance.
m Electronic mass (9.03 x 10^{-21} gram sevens).
n Repeller mode number. The number of cycles drift is $n + \frac{3}{4}$ for "optimum drift".
p Reduced power (9.5).
r Radius of grid wire, radius of tubes forming gap.
t Time, seconds.
u_0 D-C velocity of electrons.

v	Total velocity (Appendix VIII only).
v	Instantaneous gap voltage
w	Real part of frequency (12.1).
x	Coordinate along beam.
y	A rectangular coordinate normal to x.
y	Half separation of planes forming symmetrical gap.
y_e	Magnitude of small signal electronic admittance.
z	A rectangular coordinate normal to x (Appendix II).
z	A variable of integration (Appendix VI).
α	Negative coefficient of the imaginary part of frequency (12.1).
β	Modulation coefficient.
β_a	Average value of modulation coefficient.
β_0	Modulation coefficient on axis.
β_r	Modulation coefficient at radius r from axis.
β_s	Root mean squared value of modulation coefficient.
β_y	Modulation coefficient at distance y from axis.
γ	$\gamma = \omega/u_0$.
ϵ	Dielectric constant of space (8.85×10^{-14} farads/cm).
θ	Drift angle in radians.
θ_g	Gap transit angle in radians.
λ	Wavelength in centimeters.
ϕ	A phase angle.
Φ	Reduced potential (g-13).
σ	Voltage standing wave ratio.
τ	Transit time.
τ	Time constant of thermal tuner.
τ_H	Cycling time on heating.
τ_C	Cycling time on cooling.
ψ	Magnetic flux linkage.
ω	Radian frequency.

THE reflex oscillator is a form of long-transit-time tube which has distinct advantages as a low power source at high frequencies. It may be light in weight, need have no magnetic focusing field, and can be made to operate at comparatively low voltages. A single closed resonator is used, so that tuning is very simple. Because the whole resonator is at the same dc voltage, high frequency by-pass difficulties are obviated.

The frequency of oscillation can be changed by several tens of megacycles by varying the repeller voltage ("electronic tuning"). This is very advantageous when the reflex oscillator is used as a beating oscillator. The electronic tuning can be used as a vernier frequency adjustment to the manual tuning adjustment or can be used to give an all-electrical automatic frequency-control. Electronic tuning also makes reflex oscillators serve well as frequency modulated sources in low power transmitters.

The single resonator tuning property makes it possible to construct oscillators whose mechanical tuning is wholly electronically controlled. Such control is achieved by making the mechanical motion which tunes the cavity subject to the thermal expansion of an element heated by electron bombardment.

The efficiency of the reflex oscillator is generally low. The wide use of the 707B, the 723A, the 726A and subsequent Western Electric tubes shows that this defect is outweighed by the advantages already mentioned.

The first part of this paper attempts to give a broad exposition of the theory of the reflex oscillator. This theoretical material provides a background for understanding particular problems arising in reflex oscillator design and operation. The second part of the paper describes a number of typical tubes designed at the Bell Telephone Laboratories and endeavors to show the relation between theory and practice.

The theoretical work is presented first because reflex oscillators vary so widely in construction that theoretical results serve better than experimental results as a basis for generalization about their properties. While the reflex oscillator is simple in the sense that some sort of theory about it can be worked out, in practice there are many phenomena which are not included in such a theory. This leaves one in some doubt as to how well any simplified theory should apply. Multiple transits of electrons, different drift times for different electron paths and space charge in the repeller region are some factors not ordinarily taken into account which, nevertheless, can be quite important. There are other effects which are difficult to evaluate, such as distribution of current density in the beam, loss of electrons on grids or on the edges of apertures and dynamic focusing. If we could provide a theory including all such known effects, we would have a tremendous number of more or less adjustable constants, and it would not be hard to fit a large body of data to such a theory, correct or incorrect.

At present it appears that the theory of reflex oscillators is important in that it gives a semi-quantitative insight into the behavior of reflex oscillators and a guide to their design. The extent to which the present theory or an extended theory will fit actual data in all respects remains to be seen.

The writers thus regard the theory presented here as a guide in evaluating the capabilities of reflex oscillators, in designing such oscillators and in understanding the properties of such tubes as are described in the second part of this paper, rather than as an accurate quantitative tool. Therefore, the exposition consists of a description of the theory of the reflex oscillator and some simple calculations concerning it, with the more complicated mathematical work relegated to a series of chapters called appendices. It is hoped that this so organizes the mathematical work as to make it assimilable and useful, and at the same time enables the casual reader to obtain a clear idea of the scope of the theory.

I. Introduction

An idealized reflex oscillator is shown schematically in Fig. 1. It has, of course, a resonant circuit or "resonator."[1] This may consist of a pair of grids forming the "capacitance" of the circuit and a single turn toroidal

[1] For a discussion of resonators, see Appendix I. It is suggested that the reader consult this before continuing with the main work in order to obtain an understanding of the circuit philosophy used and a knowledge of the symbols employed.

coil forming the "inductance" of the circuit. Such a resonator behaves just as do other resonant circuits. Power may be derived from it by means of a coupling loop linking the magnetic field of the single turn coil. An electron stream of uniform current density leaves the cathode and is shot across the "gap" between the two grids, traversing the radio-frequency field in this gap in a fraction of a cycle. In crossing the gap the electron stream is velocity modulated; that is, electrons crossing at different times gain

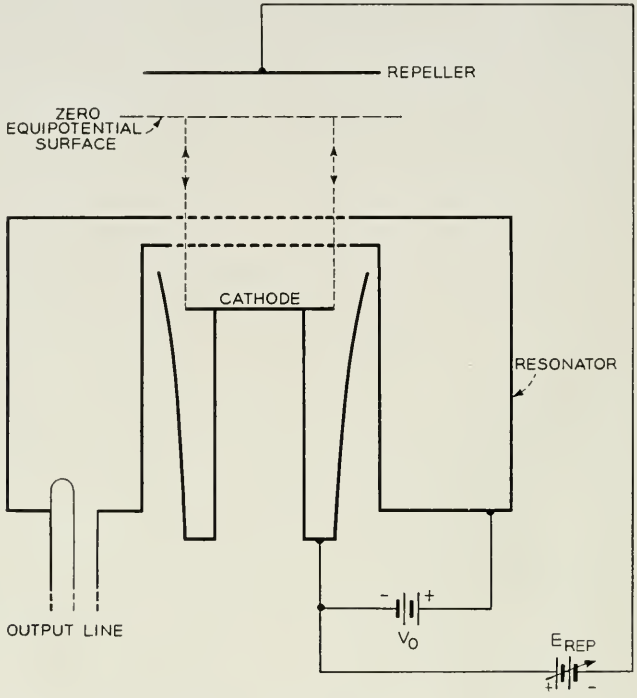


Fig. 1.—An idealized reflex oscillator with grids, shown in cross-section.

different amounts of kinetic energy from the radio-frequency voltage across the gap.[2] The velocity modulated electron stream is shot toward a negative repeller electrode which sends it back across the gap. In the "drift space" between the gap and the repeller the electron stream becomes "bunched" and the bunches of electrons passing through the radio frequency field in the gap on the return transit can give up power to the circuit if they are returned in the proper phase.

[2] The most energy any electron gains is βV electron volts, where V is the peak radio frequency voltage across the gap and β is the "modulation coefficient" or "gap factor", and is always less than unity. β depends on gap configuration and transit angle across the gap, and is discussed in Appendix II.

The vital features of the reflex oscillator are the bunching which takes place in the velocity modulated electron stream in the retarding field between the gap and the repeller and the control of the returning phase of the bunches provided by the adjustment of the repeller voltage. The analogy of Fig. 2 explains the cause of the bunching. The retarding drift field may

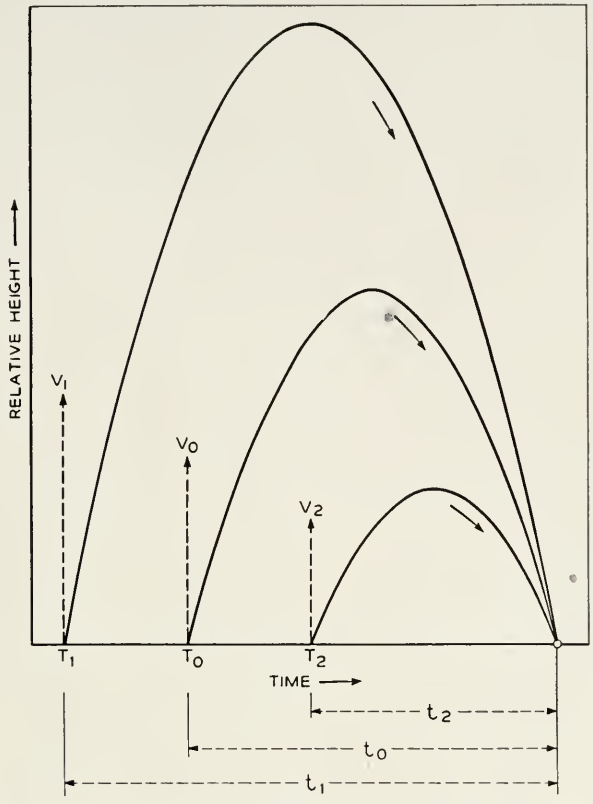


Fig. 2.—The motion of electrons in the repeller space of a reflex oscillator may be likened to that of balls thrown upward at different times. In this figure, height is plotted vs time. If a ball is thrown upward with a large velocity of V_1 at a time T_1, another with a smaller velocity at a later time T_0 and a third with a still smaller velocity at a still later time T_2 the three balls can be made to fall back to the initial level at the same time.

be likened to the gravitational field of the earth. The drift time is analogous to the time a ball thrown upwards takes to return. If the ball is thrown upward with some medium speed v_0, it will return in some time t_0. If it is thrown upward with a low speed v_1 smaller than v_0, the ball will return in some time t_1 smaller than t_0. If the ball is thrown up with a speed v_2 greater than v_0, it returns in some time t_2 greater than t_0. Now imagine three balls thrown upward in succession, evenly spaced but with large,

medium, and small velocities, respectively.[3] As the ball first thrown up takes a longer time to return than the second, and the third takes a shorter time to return than the second, when the balls return the time intervals between arrivals will be less than between their departures. Thus time-position "bunching" occurs when the projection velocity with which a uniform stream of particles enters a retarding field is progressively decreased.

Figure 3 demonstrates such bunching as it actually takes place in the retarding field of a reflex oscillator. An electron crossing the gap at phase A

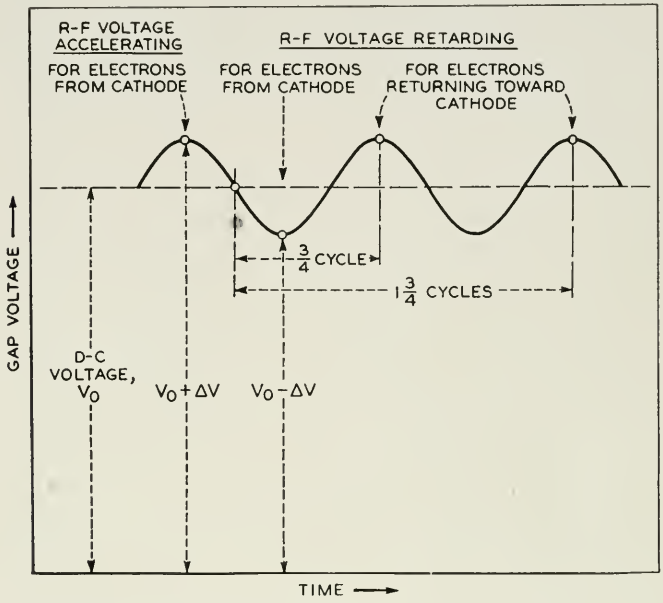


Fig. 3.—The drift time for transfer of energy from the bunched electron stream to the resonator can be deduced from a plot of gap voltage vs time.

is equivalent to the first ball since its velocity suffers a maximum increase, an electron crossing at phase B corresponds to the ball of velocity v_0 where for the electron V_0 corresponds to the d.c. injection velocity, and finally an electron crossing at phase C corresponds to the third ball since it has suffered a maximum decrease in its velocity. The electrons tend to bunch about the electron crossing at phase B. To a first order in this process no energy is taken from the cavity since as many electrons give up energy as absorb it.

The next step in the process is to bring back the grouped electrons in such a phase that they give the maximum energy to the r.f. field. Now, $\frac{3}{4}$ of a cycle after the gap voltage in a reflex oscillator such as that shown in Fig. 1 is changing most rapidly from accelerating to retarding for electrons

[3] The reader is not advised to try this experimentally unless he has juggling experience.

going from the cathode, it has a maximum retarding value for electrons returning through the gap. This is true also for $1\frac{3}{4}$ cycles, $2\frac{3}{4}$ cycles, $n + \frac{3}{4}$ cycles. Hence as Fig. 3 shows if the time electrons spend in the drift space is $n + \frac{3}{4}$ cycles, the electron bunches will return at such time as to give up energy to the resonator most effectively.

II. Electronic Admittance—Simple Theory

In Appendix III an approximate calculation is made of the fundamental component of the current in the electron stream returning through the gap of a reflex oscillator when the current is caused by velocity modulation and drift action in a uniform retarding field. The restrictive assumptions are as follows:

(1) The radio-frequency voltage across the gap is a small fraction of the d-c accelerating voltage.

(2) Space charge is neglected. Amongst other things this assumes no interaction between incoming and outgoing streams and is probably the most serious departure from the actual state of affairs.

(3) Variations of modulation coefficient for various electron paths are neglected.

(4) All sidewise deflections are neglected.

(5) Thermal velocities are neglected.

(6) The electron flow is treated as a uniform distribution of charge.

(7) Only two gap transits are allowed.

An expression for the current induced in the circuit (β times the electron convection current) is

$$i = 2I_0\beta J_1\left(\frac{\beta V\theta}{2V_0}\right)e^{j(\omega t-\theta)}. \tag{2.1}$$

Here the current is taken as positive if the beam in its second transit across the gap absorbs energy from the resonator. The voltage across the gap at the time the stream returns referred to the same phase reference as the current is $v = Ve^{-j(\omega t-\pi/2)}$. Hence the admittance appearing in shunt with the gaps will be

$$Y_e = \frac{2I_0\beta}{V} J_1\left(\frac{\beta V\theta}{2V_0}\right)e^{j((\pi/2)-\theta)}. \tag{2.2}$$

For small values of V approaching zero this becomes

$$Y_{es} = \frac{I_0\beta^2\theta}{2V_0}e^{j((\pi/2)-\theta)} = y_e e^{j((\pi/2)-\theta)} \tag{2.3}$$

$$y_e = \frac{I_0\beta^2\theta}{2V_0}. \tag{2.4}$$

Here we define Y_{es} as the small signal value of the admittance, and y_e as the magnitude of this quantity. If we plot the function Y_{es} in a complex admittance plane it takes the form of a geometric spiral as shown in Fig. 4.

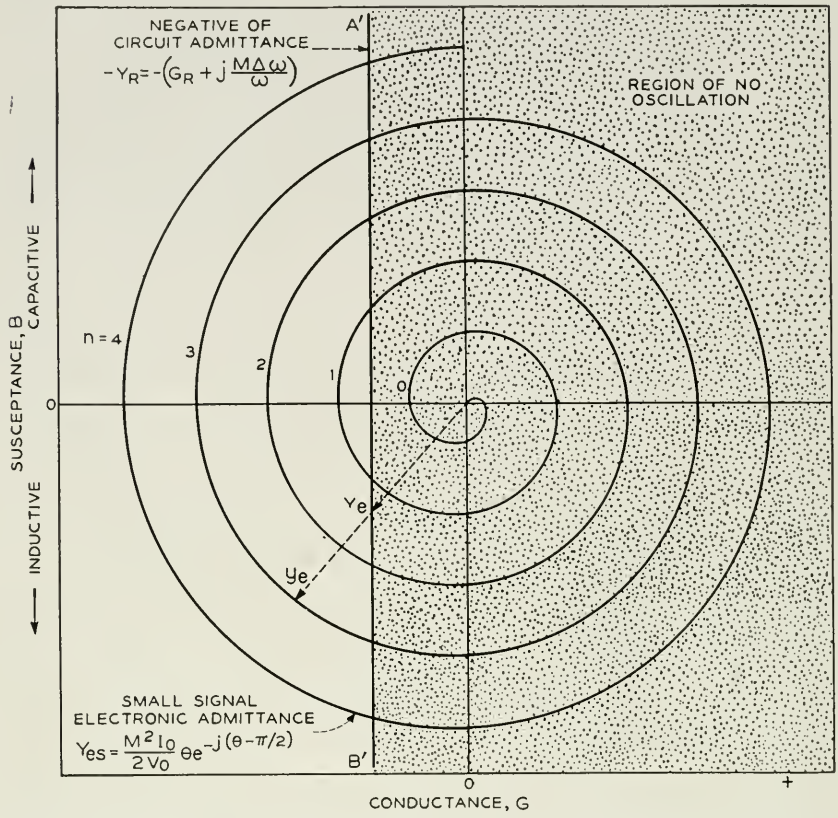


Fig. 4.—The negative of the circuit admittance (the heavy vertical line) and the small signal electronic admittance (the spiral) are shown in a plot of susceptance vs conductance. Each position along the circuit admittance line corresponds to a certain frequency. Each position along the spiral corresponds to a certain drift angle.

Such a presentation is very helpful in acquiring a qualitative understanding of the operation of a reflex oscillator.

In Appendix I it is shown that the admittance of the resonant circuit in the neighborhood of resonance is very nearly

$$Y_R = G_R + j2M\Delta\omega/\omega \tag{2.5}$$

where G_R is a constant. The negative of such an admittance has been plotted in Fig. 4 as the vertical line $A'B'$. Vertical position on this line is

proportional to the frequency at which the resonator is driven. The condition for stable oscillation is

$$Y_R + Y_e = 0. \tag{2.6}$$

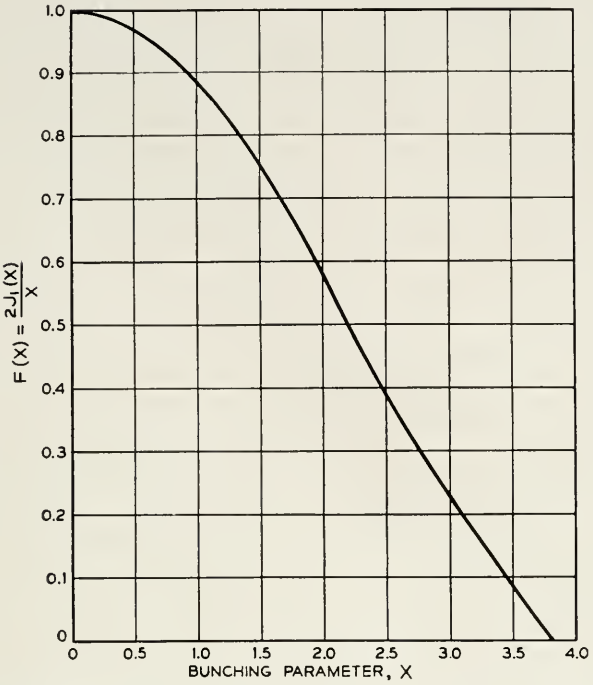


Fig. 5.—Relative amplitude of electronic admittance vs the bunching parameter X The bunching parameter increases linearly with radio frequency gap voltage so that this curve shows the reduction in magnitude of electronic admittance with increasing voltage.

We may rewrite (2.2) for any given value of θ as

$$Y_e = y_e F(V) e^{j((\pi/2)-\theta)} \tag{2.7}$$

where

$$F(V) = \frac{2J_1\dfrac{(\beta V\theta)}{(2V_0)}}{\dfrac{\beta V\theta}{2V_0}} = \frac{2J_1(X)}{X}. \tag{2.8}$$

The quantity

$$X = \frac{\beta V\theta}{2V_0} \tag{2.9}$$

is called the bunching parameter. A plot of the function $F(V)$ vs X is shown in Fig. 5. For any given value of θ and for fixed operating conditions

it is a function of V only and its action is clearly to reduce the small signal value of the admittance until condition (2.6) is satisfied. It will be observed that this function affects the magnitude only and not the phase of the admittance.

Thus, as indicated in Fig. 4, when oscillation starts the admittance is given by the radius vector of magnitude y_e, terminating on the spiral, and as the oscillation builds up this vector shrinks until in accordance with (12.6) it terminates on the circuit-admittance line $A'B'$, which is the locus of vectors $(-Y_R)$. The electronic admittance vector may be rotated by a change in the repeller voltage which changes the value of θ. This changes the vertical intercept on line $A'B'$, and since the imaginary component of the circuit admittance, that is the height along $A'B'$, is proportional to frequency, this means that the frequency of oscillation changes. It is this property which is known as electronic tuning.

Oscillation will cease when the admittance vector has rotated to an angle such that it terminates on the intersection of the spiral and the circuit-admittance line $A'B'$. It will be observed that the greater is the number of cycles of drift the greater is the electronic tuning to extinction. While it is not as apparent from this diagram, it is also true that the greater the number of cycles of drift the greater the electronic tuning to intermediate power points. Vertical lines farther to the left correspond to heavier loads, and from this it is apparent that the electronic tuning to extinction decreases with the loading. By sufficient loading it is possible to prevent some repeller modes (i.e. oscillations of some n values) from occurring. Since losses in the resonant cavity of the oscillator represent some loading, some modes of low n value will not occur even in the absence of external loading.

III. Power Production for Drift Angle of $(n + \frac{3}{4})$ Cycles

Now, from equation (2.2) it may be seen that Y_e will be real and negative for $\theta = \theta_n = (n + \frac{3}{4})2\pi$. Because θ also appears in the argument of the Bessel function this value of θ is not exactly the value to make the real component of Y_e a maximum. However, for the reasonably large values of n encountered in practical oscillators this is a justifiable approximation. Suppose, then, we consider the case of $n + \frac{3}{4}$ cycles drift, calling this an optimum drift time. Using the value of n as a parameter we plot the magnitude of the radio-frequency electron current in the electron stream returning across the gap given by equation (2.1) as a function of the radio-frequency voltage across the gap. This variation is shown in Fig. 6. As might be expected, the greater the number of cycles the electrons drift in the drift space, the lower is the radio-frequency gap voltage required to produce a given amount of bunching and hence a given radio frequency electron current. It may be seen from Fig. 6 that as the radio-frequency gap voltage

is increased, the radio-frequency electron current gradually increases until a maximum value is reached, representing as complete bunching as is possible, after which the current decreases with increasing gap voltage. The maximum value of the current is approximately the same for various drift times, but occurs at smaller gap voltages for longer drift times.

The radio-frequency power produced is the voltage times the current. As the given maximum current occurs at higher voltages for shorter drift

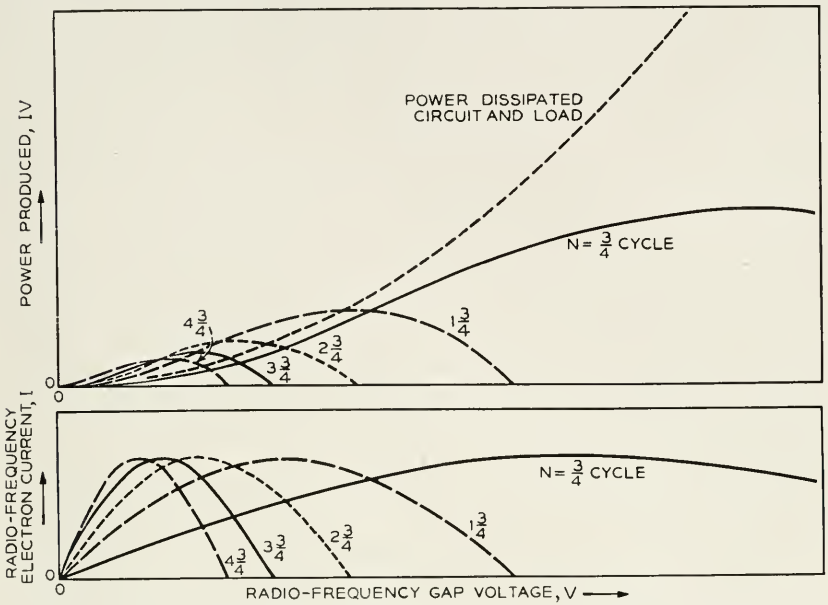


Fig. 6.—Radio frequency electron convection current I and the radio frequency power given up by the electron stream can be plotted vs the radio frequency gap voltage V for various drift times measured in cycles. Maximum current occurs at higher voltage for shorter drift times. For a given number of cycles drift, maximum power occurs at a higher gap voltage than that for maximum current. If the power produced for a given drift time is higher at low voltages than the power dissipated in the circuit and load (dashed curve), the tube will oscillate and the amplitude will adjust itself to the point at which the power dissipation and the power production curves cross.

times, the maximum power produced will be greater for shorter drift times. This is clearly brought out in the plots of power vs. voltage shown in Fig. 6.

The power dissipated in the circuits and load will vary as the square of the radio-frequency voltage. Part of this power will go into the load coupled with the circuit and part into unavoidable circuit losses. A typical curve of power into the circuit and load vs. radio-frequency voltage is shown in Fig. 6. Steady oscillation will take place at the voltage for which the power production curve crosses the power dissipation curve. For instance, in Fig. 6 the power dissipation curve crosses the power production curve for

$1\frac{3}{4}$ cycles drift at the maximum or hump of the curve. This means that the circuit impedance for the dissipation curve shown is such as to result in maximum production of power for $1\frac{3}{4}$ cycles drift. For $2\frac{3}{4}$ cycles drift and for longer drifts, the power dissipation curve crosses the power production curves to the right of the maximum and hence the particular circuit loading shown does not result in maximum power production for these longer drift times. This is an example of operation with lighter than optimum load. The power dissipation curve might cross the power production curve to the left of the maximum, representing a condition of too heavy loading for production of maximum power output. The power dissipation curve in Fig. 6 lies always above the power production curve for a drift of $\frac{3}{4}$ cycles. This means that the oscillator for which the curves are drawn, if loaded to give the power dissipation curve shown, would not oscillate with the short drift time of $\frac{3}{4}$ cycles, corresponding to a very negative repeller voltage.

In general, the conclusions reached by examining Fig. 6 are borne out in practice. The longer the drift time, that is, the less negative the repeller, the lower is the power output. For very negative repeller voltages, however, corresponding to very short drift times, the power either falls off, which means that most of the available power is dissipated in circuit losses, or the oscillator fails to operate at all because, for all gap voltages, the power dissipated in circuit losses is greater than the power produced by the electron stream.

Having examined the situation qualitatively, we want to make a somewhat more quantitative investigation, and to take some account of circuit losses. In the course of this we will find two parameters are very important. One is the parameter X previously defined by equation (2.9), which expresses the amount of bunching the beam has undergone. In considering a given tube with a given drift time, the important thing to remember about X is that it is proportional to the *r-f* gap voltage V. For $\theta = \theta_n$ expression (2.2) is a pure conductance and we can express the power produced by the electron stream as one half the square of the peak *r-f* voltage times the circuit conductance which for stable oscillation is equal to the negative of the electronic conductance given by (2.2). This may be written with the aid of (2.9) as

$$P = 2(I_0 V_0/\theta_n) X J_1(X). \tag{3.1}$$

Suppose we take into account the resonator losses but not the power lost in the output circuit, which in a well designed oscillator should be small. If the resonator has a shunt resonant conductance (including electronic loading) of G_R, the power dissipated in the resonator is

$$P_R = V^2 G_R/2. \tag{3.2}$$

Then the power output for θ_n is

$$P = 2(I_0V_0/\theta_n)XJ_1(X) - V^2G_R/2. \tag{3.3}$$

The efficiency, η, is given by

$$\eta = \frac{P}{P_{DC}} = \frac{2}{\theta}\left[XJ_1(X) - \frac{G_R\theta_n V^2}{4I_0V_0}\right]. \tag{3.4}$$

From (2.4) and (2.9)

$$\frac{\theta_n V^2}{2I_0V_0} = \frac{1}{y_e}X^2. \tag{3.5}$$

Hence we may write

$$\eta = \frac{1}{(n + 3/4)}\frac{1}{\pi}\left[XJ_1(X) - \frac{G_R}{y_e}\frac{X^2}{2}\right]. \tag{3.6}$$

Let us write $\eta = \dfrac{H}{N}$ where $N = (n + \frac{3}{4})$. We may now make a generalized examination of the effect of losses on the efficiency by examining the function

$$H = (1/\pi)[XJ_1(X) - (G_R/y_e)X^2/2]. \tag{3.7}$$

Thus, the efficiency for $\theta = \theta_n$ is inversely proportional to the number of cycles drift and is propotional to a factor H which is a function of X and of the ratio G_R/y_e, that is, the ratio of resonator loss conductance to small signal electronic conductance.[4] For $n + \frac{3}{4}$ cycles drift, the small signal electronic conductance is equal to the small signal electronic admittance.

For a given value of G_R/y_e there is an optimum value of X for which H has a maximum H_m. We can obtain this by differentiating (3.7) with respect to X and setting the derivative equal to zero, giving

$$\begin{aligned} XJ_0(X) - (G_R/y_e)X &= 0 \\ J_0(X) &= (G_R/y_e). \end{aligned} \tag{3.8}$$

If we put values from this into (3.7) we can obtain H_m as a function of G_R/y_e. This is plotted in Fig. 7. The considerable loss of efficiency for values of G_R/y_e as low as .1 or .2 is noteworthy. It is also interesting to note that for G_R/y_e equal to $\frac{1}{3}$, the fractional change in power is equal to the fractional change in resonator resistance, and for G_R/y_e greater than $\frac{1}{3}$, the fractional power change is greater than the fractional change in resonator resistance. This helps to explain the fall in power after turn-on in some tubes, for an increase in temperature can increase resonator resistance considerably.

[4] An electronic damping term discussed in Appendix VIII should be included in resonator losses. The electrical loss in grids is discussed briefly in Appendix IX.

In the expression for the admittance, the drift angle, θ, appears as a factor. This factor plays a double role in that it determines the phase of the admittance but also in a completely independent manner it determines, in part, the magnitude of the admittance. θ as it has appeared in the foregoing analysis, which was developed on the basis of a linear retarding field, is the actual drift angle in radians. As will be shown in a later section, certain special repeller fields may give effective drift action for a given angle greater than the same angle in a linear field. Such values of effective drift angle may have fractional optimum values although the phase must still be such as to give within the approximations we have been using a pure conductance at optimum. In order to generalize the following work we will speak of an effective drift time in cycles, $N_E = FN$, where N is the actual drift time in cycles, $n + \frac{3}{4}$, and F is the number of times this drift is more effective than the drift in a linear field.

Suppose we have a tube of given β^2, I_0, V_0 and resonator loss G_R and wish to find the optimum effective drift time, FN, and determine the effect on the efficiency of varying FN. It will be recalled that for very low losses we may expect more power output the fewer the number of cycles drift. However the resonator losses may cut heavily into the generated power, for short drift angles. With short drift angles the optimum load conductance becomes small compared to the loss conductance so that although the generated power is high only a small fraction goes to the useful load. There is, therefore, an optimum value which can be obtained using the data of Fig. 7. We define a parameter

$$K = \frac{2V_0}{\beta^2 I_0} G_R \tag{3.9}$$

which compares the resonator loss conductance with the small signal electronic admittance per radian of driftan gle. Then in terms of K, $\frac{G_R}{y_e} = \frac{K}{\theta}$. Hence, for a fixed value of K, various values of $\theta = 2\pi FN$ define values of $\frac{G_R}{y_e}$. When one uses these values in connection with Fig. 7 he determines the corresponding values of H_m and hence the efficiency, $\eta = \frac{H_m}{FN}$. These values of η are plotted against FN as in Fig. 8 with values of K as a parameter. In this plot K is a measure of the lossiness of the tube. The optimum drift angle for any degree of lossiness is evident as the maximum of one of these curves.

The maximum power outputs in various repeller modes, $n = 0$, 1 etc. and the repeller voltages for these various power outputs correspond to discrete values of n and FN lying along a curve for a particular value of K.

Thus, the curves illustrate the variation of power from mode to mode as the repeller voltage is changed over a wide range.

Changes in resonator loss or differences in loss between individual tubes of the same type correspond to passing from a curve for one value of K to a curve for another value of K.

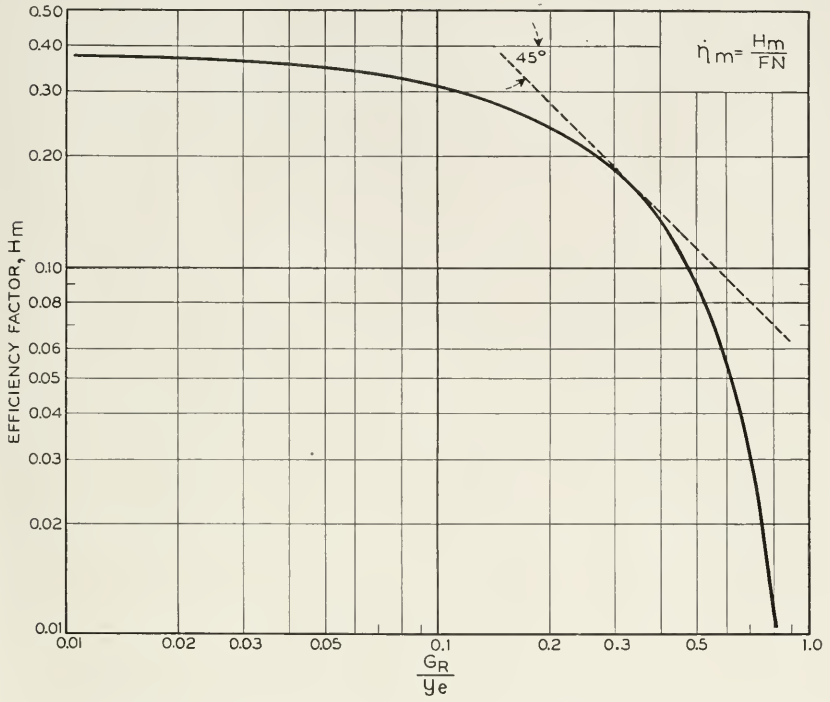


Fig. 7.—Efficiency factor H_m vs the ratio of resonator loss conductance to the small signal electronic admittance. Efficiency changes rapidly with load as the loss conductance approaches in magnitude the small signal electronic admittance. The efficiency is inversely proportional to the number of cycles drift.

It will be observed from this that although, from an efficiency standpoint, it is desirable to work at low values of drift time such low drift times lead to an output strongly sensitive to changes in resonator losses.

Perhaps the most important question which the user of the oscillator may ask with regard to power production for optimum drift is; what effect does the external load have upon the performance? If we couple lightly to the oscillator the *r-f* voltage generated will be high but the power will not be extracted. If we couple too heavily the voltage will be low, the beam will not be efficiently modulated and the power output will be low. There is

apparently an optimum loading. Best output is not obtained when the external load matches the generator impedance as in the case of an amplifier.

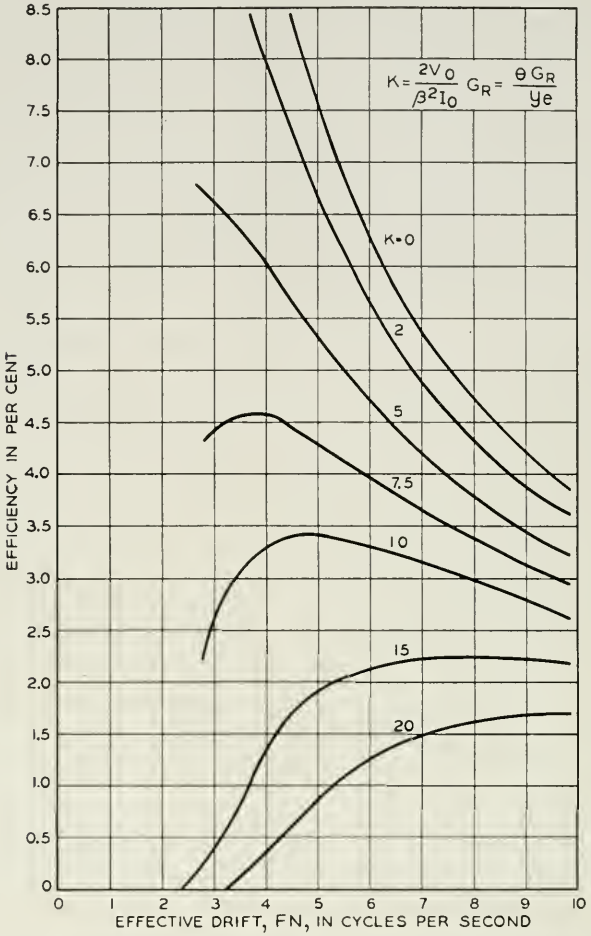


Fig. 8.—Efficiency in per cent vs the effective cycles drift for various values of a parameter K which is proportional to resonator loss. These curves indicate how the power output differs for various repeller modes for a given loss. Optimum power operating points will be represented by points along one of these curves. For a very low loss resonator, the power is highest for short drift times and decreases rapidly for higher repeller modes. Where there is more loss, the power varies less rapidly from mode to mode.

We return to equation (3.7) for H and assume that we are given various values of $\frac{G_R}{y_e}$. With these values as parameters we ask what variation in efficiency may be expected as we vary the ratio of the load conductance, G_L, to the small signal admittance, y_e. When $\frac{G_L}{y_e}+\frac{G_R}{y_e}=1$ oscillation

will just start and no power output will be obtained. We can state the general condition for stable operation as $Y_e + Y_C = 0$, where Y_C is the vector sum of the load and circuit admittances. For the optimum drift time this becomes

$$\frac{G_C}{y_e} = \frac{2J_1(X)}{X} \tag{3.10}$$

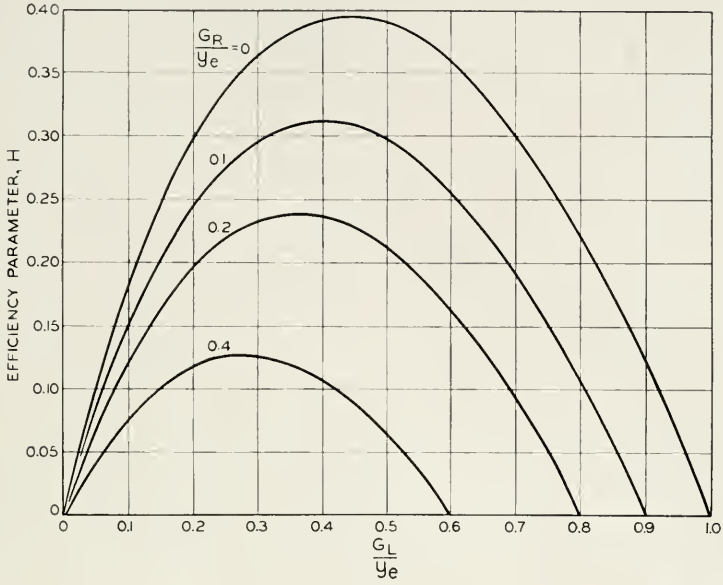


Fig. 9.—Efficiency parameter H vs the ratio of load conductance to the magnitude of the small signal electronic admittance. Curves are for various ratios of resonator loss conductance to small signal electronic admittance. The curves are of similar shapes and indicate that the tube will cease oscillating ($H = 0$) when loaded by a conductance about ½ ice as large as that for optimum power.
t

where

$$\frac{G_C}{y_e} = \frac{G_L + G_R}{y_e}. \tag{3.11}$$

Hence for a given value of $\frac{G_R}{y_e}$ we may assume values for $\frac{G_L}{y_e}$ between zero and $1 - \frac{G_R}{y_e}$ and these in (3.11) will define values of X. These values of X substituted in (3.7) will define values of H which we then plot against the assumed values for $\frac{G_L}{y_e}$, as in Fig. 9. Thus we have the desired function of the variation of efficiency factor against load.

From the curves of Fig. 9 it can be seen that the maximum efficiency is obtained when the external conductance is made equal to approximately half the available small signal conductance; i.e. $\frac{1}{2}(y_e - G_R)$. This can be seen more clearly in the plot of Fig. 10. Equation (3.8) gives the condition for maximum efficiency as

$$\frac{G_R}{y_e} = J_0(X).$$

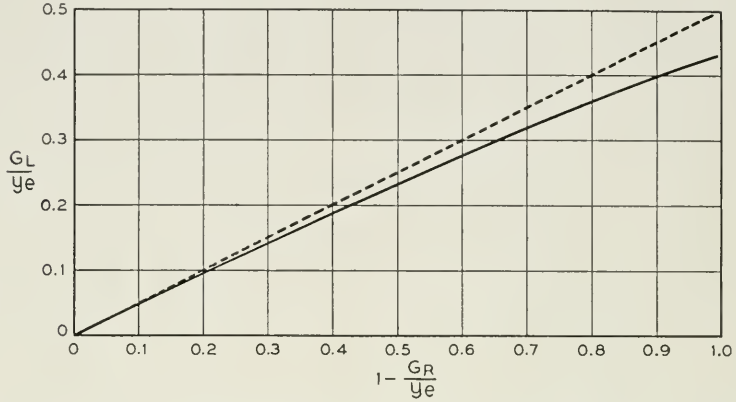


Fig. 10.—The abscissa measures the fractional excess of electronic negative conductance over resonator loss conductance. The ordinate is the load conductance as a fraction of electronic negative conductance. The tube will go out of oscillation for a load conductance such that the ordinate is equal to the abscissa. The load conductance for optimum power output is given by the solid line. The dashed line represents a load conductance half as great as that required to stop oscillation.

If we assume various values for $\frac{G_R}{y_e}$ these define values of X_0 which when substituted in

$$\frac{G_L}{y_e} = \frac{G_C}{y_e} - \frac{G_R}{y_e} = \frac{2J_1(X_0)}{X_0} - J_0(X_0) \tag{3.12}$$

give the value of the external load for optimum power. We plot these data against the available conductance

$$\left(1 - \frac{G_R}{y_e}\right) = 1 - J_0(X) \tag{3.13}$$

as shown in Fig. 10.

In Fig. 10 there is also shown a line through the origin of slope 1/2. It can thus be seen that the optimum load conductance is slightly less than half the available small signal or starting conductance. This relation is independent of the repeller mode, i.e. of the value FN. This does not mean

that the load conductance is independent of the mode, since we have expressed all our conductances in terms of y_e, the small signal conductance, and this of course depends on the mode. What it does say is that, regardless of the mode, if the generator is coupled to the load conductance for maximum output, then, if that conductance is slightly more than doubled oscillation will stop. It is this fact which should be borne in mind by the circuit designer. If greater margin of safety against "pull out" is desired it can be obtained only at the sacrifice of efficiency.

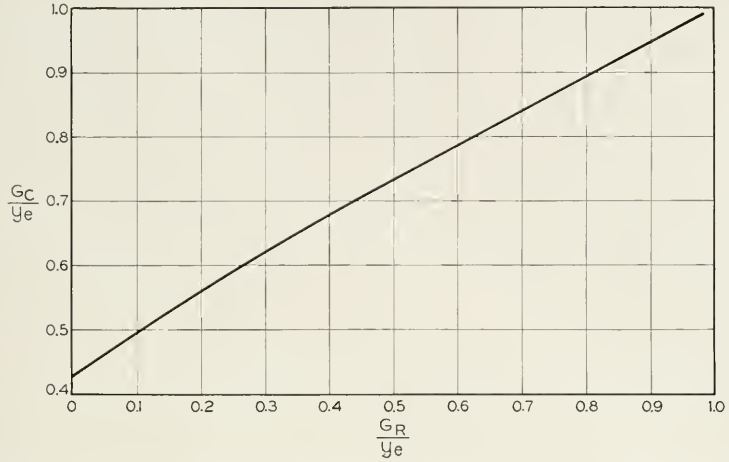


Fig. 11.—The ratio of total circuit conductance for optimum power to small signal electronic admittance, vs the ratio of resonator loss conductance to small signal electronic admittance.

An equivalent plot for the data of Fig. 10, which will be of later use, is shown in Fig. 11. This gives the value of $\frac{G_C}{y_e}$ for best output for various values of $\frac{G_R}{y_e}$.

IV. Effect of Approximations

The analysis presented in Section II is misleading in some respects. For instance, for a lossless resonator and $N = \frac{3}{4}$ cycles, the predicted efficiency is 53%. However, our simple theory tells us that to get this efficiency, the radio-frequency gap voltage V multiplied by the modulation coefficient β (that is, the energy change an electron suffers in passing the gap) is $1.018V_0$. This means that (a) some electrons would be stopped and would not pass the gap (b) many other electrons would not be able to pass the gap against a retarding field after returning from drift region (c) some electrons would

cross the gap so slowly that for them β would be very small and their effect on the circuit would also be small (d) there might be considerable loading of the resonator due to transit time effects in the gap. Of course, it is not justifiable to apply the small signal theory in any event, since it was derived on the assumption that βV is small compared with V_0.

In Appendix IV there is presented a treatment by R. M. Ryder of these Laboratories in which it is not assumed that $\beta V \ll V_0$. This work does not, however, take into account variation of β with electron speed or the possibility of electrons being turned back at the gap.

For drift angles of $1\frac{3}{4}$ cycles and greater, the results of Ryder's analysis are almost indistinguishable from those given by the simple theory, as may be seen by examining Figs. 128–135 of the Appendix. His curves approach the curves given by the simple theory for large values of n.

For small values of n, and particularly for $\frac{3}{4}$ cycles drift, Ryder's work shows that optimum power is obtained with a drift angle somewhat different from $n + \frac{3}{4}$ cycles. Also, Fig. 131 shows that the phase of the electronic admittance actually varies somewhat with amplitude, and Fig. 130 shows that its magnitude does not actually pass through zero as the amplitude is increased.

The reader is also referred to a paper by A. E. Harrison.[5]

The reader may feel at this point somewhat uneasy about application of the theory to practice. In most practical reflex oscillators, however, the value of n is 2 or greater, so that the theory should apply fairly well. There are, however, so many accidental variables in practical tubes that it is well to reiterate that the theory serves primarily as a guide, and one should not expect quantitative agreement between experiment and theory. This will be apparent in later sections, where in a few instances the writers have made quantitative calculations.

V. Special Drift Fields

In the foregoing sections a theory for a reflex oscillator has been developed on the assumption that the repeller field is a uniform retarding electrostatic field. Such a situation rarely occurs in practice, partly because of the difficulty of achieving such a field and partly because such a field may not return the electron stream in the manner desired. In an effort to get some information concerning actual drift fields, we may extend the simple theory already presented to include such fields by redefining X as

$$X = \beta V F \theta / 2V_0. \tag{5.1}$$

Here the factor F is included. As defined in Section *III* this is the factor which relates the effectiveness of a given drift field in bunching a velocity

[5] A. E. Harrison, "Graphical Methods for Analysis of Velocity Modulation Bunching," *Proc. I.R.E.*, 33.1, pp. 20–32, June 1945.

modulated electron stream with the bunching effectiveness of a field with the same drift angle θ but with a linear variation of potential with distance. Suppose, for instance, that the variation of transit time, τ, with energy gained in crossing the gap V is for a given field

$$\partial\tau/\partial V \tag{5.2}$$

and for a linear potential variation and the same drift angle

$$(\partial\tau/\partial V)_1. \tag{5.3}$$

Then the factor F is defined as

$$F = (\partial\tau/\partial V)/(\partial\tau/\partial V)_1. \tag{5.4}$$

In appendix V, F is evaluated in terms of the variation of potential with distance.

The efficiency is dependent on the effectiveness of the drift action rather than on the total number of cycles drift except of course for the phase requirements. Thus, for a nonlinear potential variation in the drift space we should have instead of (3.7)

$$\eta = H/FN. \tag{5.5}$$

In the investigation of drift action, one procedure is to assume a given drift field and try to evaluate the drift action. Another is to try to find a field which will produce some desirable kind of drift action. As a matter of fact, it is easy to find the best possible drift field (from the point of view of efficiency) under certain assumptions.

The derivation of the optimum drift field, which is given in appendix VII, hinges on the fact that the time an electron takes to return depends only on the speed with which it is injected into the drift field. Further, the variation in modulation coefficient for electrons returning with different speeds is neglected. With these provisos, the optimum drift field is found to be one in which electrons passing the gap when the gap voltage is decelerating take π radians to return, and electrons which pass the gap when the voltage is accelerating take 2π radians to return, as illustrated graphically in Fig. 136, Appendix VII. A graph of potential vs. distance from gap to achieve such an ideal drift action is shown in Fig. 137 and the general appearance of electrodes which would achieve such a potential distribution approximately is shown in Fig. 138.

With such an ideal drift field, the efficiency of an oscillator with a lossless resonator is

$$\eta_i = (2/\pi)(\beta V/V_0). \tag{5.6}$$

For a linear potential variation in the drift space, at the optimum *r-f* gap voltage, according to the approximate theory presented in Section III the efficiency for a lossless resonator is

$$\eta = (.520)(\beta V/V_0). \tag{5.7}$$

Comparing, we find an improvement in efficiency for the ideal drift field in the ratio

$$\eta_i/\eta = 1.23, \tag{5.8}$$

or only about 20%. Thus, the linear drift field is quite effective. The ideal drift field does have one advantage; the bunching is optimum for all gap voltages or, for a given gap voltage, for all modulation coefficients since ideally an infinitesimal *a-c* voltage will change the transit time from π to 2π and completely bunch the beam. This should tend to make the efficiency high despite variations in β over various parts of the electron flow. The limitation imposed by the fact that electrons cannot return across the gap against a high voltage if they have been slowed up in their first transit across the gap remains.

This last mentioned limitation is subject to amelioration. In one type of reflex oscillator which has been brought to our attention the electrons cross the gap the first time in a region in which the modulation coefficient is small. If the gap has mesh grids, a hole may be punched in the grids and a beam of smaller diameter than the hole focussed through it. Then the beam may be allowed to expand and recross a narrow portion of the gap, where the modulation coefficient is large. Thus, in the first crossing no electrons lose much energy (because β is small) and in the second crossing all can cross the gap where βV is large and hence can give up a large portion of their energy.

VI. Electronic Gap Loading

So far, attention has been concentrated largely on electronic phenomena in the drift or repeller region. To the long transit time across the gap there has been ascribed merely a reduction in the effect of the voltage on the electron stream by the modulation coefficient β. Actually, the long transit across the gap can give rise to other effects.

One of the most obvious of these other effects is the production of an electronic conductance across the gap. If it is positive, such a conductance acts just as does the resonator loss conductance in reducing the power output. Petrie, Strachey and Wallis of Standard Telephones and Cables have treated this matter in an interesting and rather general way. Their work, in a slightly modified form, appears in Appendix VIII, to which the reader is referred for details.

The work tells us that, considering longitudinal fields only, the electron flow produces a small signal conductance component across the gap

$$G_{eL} = \frac{I_0}{4V_0}\frac{\partial\beta^2}{\partial\gamma} \tag{6.1}$$

$$\gamma = \frac{\omega}{u_0}. \tag{6.2}$$

Here β is the modulation coefficient and u_0 is the electron speed. I_0 and V_0 are beam current and beam voltage. If the gap has a length d, the transit angle across it is $\theta_g = \gamma d$ and (6.1) may be rewritten

$$G_{eL} = -\frac{I_0\theta_g}{4V_0}\frac{\partial\beta^2}{\partial\theta_g}. \tag{6.3}$$

It is interesting to compare this conductance with the magnitude of the small-signal electronic admittance, y_e. In doing so, we should note that the current crosses the gap twice, and on each crossing produces an electronic conductance. Thus, the appropriate comparison between loss conductance and electronic admittance is $2G_{eL}/y_e$. Using (6.3) we obtain

$$2G_{eL}/y_e = -\frac{\theta_g}{\theta}\frac{1}{\beta^2}\frac{\partial\beta^2}{\partial\theta_g}. \tag{6.4}$$

Usually, the drift angle θ is much larger than the gap transit angle θ_g. Further, if we examine the curves for modulation coefficient β which are given in Appendix II, we find that $(\partial\beta^2/\partial\theta_g)/\beta^2$ will not be very large. Thus, we conclude that in general the total loss conductance for longitudinal fields will be small compared with the electronic admittance. An example in Appendix VIII gives $(2G_{eL}/y_e)$ as about 1/10. It seems that this effect will probably be less important than various errors in the theory of the reflex oscillator.

Even though this electronic gap loading is not very large, it may be interesting to consider it further. We note, for instance, that the conductance G_{eL} is positive when β^2 decreases as gap transit angle increases. For parallel fine grids this is so from $\theta_g = 0$ to $\theta_g = 2\pi$ (see Fig. 119 of Appendix II). At $\theta_g = 2\pi$, where $\beta = 0$, $\partial\beta^2/\partial\theta_g = 0$, and the gap loading is zero. In a region beyond $\theta_g = 2\pi$, $\partial\beta^2/\partial\theta_g$ becomes positive and the gap conductance is negative. Thus, for some transit angles a single gap can act to produce oscillations. For still larger values of θ_g, G_{eL} alternates between positive and negative. Gap transit angles of greater than 2π are of course of little interest in connection with reflex oscillators, as for such transit angles β is very small.

For narrow gaps with large apertures rather than fine grids, $\partial\beta^2/\partial\theta_g$

never becomes very negative and may remain positive and the gap loading conductance due to longitudinal fields be always positive. In such gaps, however, transverse fields can have important effects, and (6.3) no longer gives the total gap conductance. Transverse fields act to throw electrons approaching the gap outward or inward, into stronger or weaker longitudinal fields, and in this manner the transverse fields can either cause the electrons to give up part of their forward velocity, transferring energy to the resonator, or to pick up forward velocity, taking energy from the resonator. An analysis of the effect of transverse fields is given in Appendix VIII, and this is applied in calculating the total conductance, due both to longitudinal and to transverse fields, of a short gap between cylinders with a uniform current density over the aperture. It is found that the transverse fields contribute a minor part of the total conductance, and that this contribution may be either positive or negative, but that the total gap conductance is always positive (see Appendix VIII, Fig. 140).

The electron flow across the gap produces a susceptive component of admittance. This susceptive component is in general more difficult to calculate than the conductive component. It is not very important; it serves to affect the frequency of oscillation slightly but not nearly so much as a small change in repeller voltage.

Besides such direct gap loading, the velocity modulation and drift action within a gap of fine grids actually produce a small bunching of the electron stream. In other words, the electron stream leaving such a gap is not only velocity modulated but it has a small density modulation as well. This convection current will persist (if space-charge debunching is not serious) and, as the electrons return across the gap, it will constitute a source of electronic admittance. We find however, that in typical cases (see Appendix VIII, (h59)–(h63)), this effect is small and is almost entirely absent in gaps with coarse grids or large apertures.

Secondary electrons produced when beam electrons strike grid wires and grid frames or gap edges constitute another source of gap loading. It has been alleged that if the frames supporting the grids or the tubes forming a gap have opposed parallel surfaces of width comparable to or larger than the gap spacing, large electron currents can be produced through secondary emission, the *r-f* field driving electrons back and forth between the opposed surfaces. It would seem that this phenomenon could take place only at quite high *r-f* levels, for an electron would probably require of the order of 100 volts energy to produce more than one secondary in striking materials of which gaps are usually constructed.

VII. Electronic Tuning—Arbitrary Drift Angle

So far, the "on tune" oscillation of reflex oscillators has been considered except for a brief discussion in Section II, and we have had to deal only with

real admittances (conductances). In this section the steady state operation in the case of complex circuit and electronic admittances will be discussed. The general condition for oscillation states that, breaking the circuit at any point the sum of the admittances looking in the two directions is zero. Particularly, the electronic admittance Y_e looking from the circuit to the electron stream, must be minus the circuit admittance Y_C, looking from the electron stream to the circuit. Here electronic admittance is used in the sense of an admittance averaged over a cycle of oscillation and fulfilling the above condition.

It is particularly useful to consider the junction of the electron stream and the circuit because the electronic admittance Y_e and the circuit admittance Y_C have very different properties, and if conditions are considered elsewhere these properties are somewhat mixed and full advantage cannot be taken of their difference.

The average electronic admittance with which we are concerned is a function chiefly of the amplitude of oscillation. Usually its magnitude decreases with increasing amplitude of oscillation, and its phase may vary as well, although this is a large signal effect not shown by the simple theory. In reflex oscillators the phase may be controlled by changing the repeller voltage. The phase and magnitude of the electronic admittance also vary with frequency. Usually, however, the rate of change with frequency is slow compared with that of the circuit admittance in the vicinity of any one resonant mode. By neglecting this change of electronic admittance with frequency in the following work, and concentrating our attention on the variation with amplitude and repeller voltage, we will emphasize the important aspects without serious error. However, the variation of electronic admittance with frequency should be kept in mind in considering behavior over frequency ranges of several per cent.[6]

The circuit admittance is, of course, independent of amplitude and is a rapidly varying function of frequency. It is partly dependent on what is commonly thought of as the resonator or resonant circuit of the oscillator, but is also profoundly affected by the load, which of course forms a part of the circuit seen from the electron stream. The behavior of the oscillator is determined, then, by the electronic admittance, the resonant circuit and the load. The behavior due to circuit and load effects applies generally to all oscillators, and the simplicity of behavior of the electronic admittance is such that similarities of behavior are far more striking than differences.

We have seen from Appendix I that at a frequency $\Delta\omega$ away from the resonant frequency ω_0 where $\Delta\omega \ll \omega_0$, the admittance at the gap may be expressed as:

$$Y_C = G_C + j2M\Delta\omega/\omega_0. \tag{7.1}$$

[6] Appendix IV discusses the variation of phase with frequency and repeller voltage. The variation of phase of electronic admittance with frequency is included in Section IX A.

Here the quantity M is the characteristic admittance of the resonator, which depends on resonator shape and is unaffected by scaling from one frequency to another. G_c is the shunt conductance due to circuit and to load. Y_c as given by (7.1) represents to the degree of approximation required the admittance of any resonant circuit and load with only one resonance near the frequency of oscillation.

It is profitable to consider again in more detail a complex admittance plot similar to Fig. 4. In Fig. 12 the straight vertical line is a plot of (7.1).

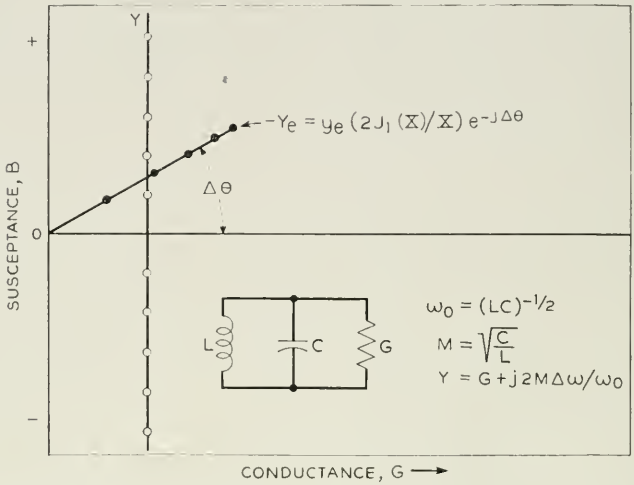


Fig. 12.—The resonator and its load can be represented as a shunt resonant circuit with a shunt conductance G. For frequencies near resonance, the conductance is nearly constant and the susceptance B is proportional to frequency, so that when susceptance is plotted vs conductance, the admittance Y is a vertical straight line. The circles mark off equal increments of frequency. The electronic admittance is little affected by frequency but much affected by amplitude. The negative of an electronic admittance Y_e having a constant phase angle $\Delta\theta$ is shown in the figure. The dots mark off equal amplitude steps. Oscillation will occur at a frequency and amplitude specified by the intersection of the curves Y and $-Y_e$.

The circles mark equal frequency increments. Now if we neglect the variation of the electronic admittance with phase, then the negative of the small signal electronic admittance on this same plot will be a vector, the locus of whose termination will be a circle. The vector is shown in Fig. 12. The dots mark off admittance values corresponding to equal amplitude increments as determined by the data of Fig. 5.

Steady oscillation will take place at the frequency and amplitude represented by the intersection of the two curves. If the phase angle $\Delta\theta$ of the $-Y_e$ curve is varied by varying the repeller voltage, the point of intersection will shift on both the Y_c curve and the $-Y_e$ curve. The shift along the

V_c curve represents a change in frequency of oscillation; the shift along the $-V_e$ curve represents a change in the amplitude of oscillation. If we know the variation of amplitude with position along the $-V_e$ curve, and the variation of frequency with position along the V_c curve, we can obtain both the amplitude and frequency of oscillation as a function of the phase of $-V_e$, which is in turn a function of repeller voltage.

From (2.3) and (2.7) we can write $-V_e$ in terms of the deviation of drift angle $\Delta\theta$ from $n + \frac{3}{4}$ cycles.

$$-V_e = y_e(2J_1(X)/X)e^{j\Delta\theta}. \tag{7.2}$$

The equation relating frequency and $\Delta\theta$ can be written immediately from inspection of Fig. 12.

$$\begin{aligned} 2M\Delta\omega/\omega_0 &= -G_C \tan \Delta\theta \\ \Delta\omega/\omega_0 &= -(G_C/2M) \tan \Delta\theta \\ \Delta\omega/\omega_0 &= -(1/2Q) \tan \Delta\theta. \end{aligned} \tag{7.3}$$

Here Q is the loaded Q of the circuit.

The maximum value of $\Delta\theta$ for which oscillation can occur (at zero amplitude) is an important quantity. From Fig. 12 this value, called $\Delta\theta_0$, is obviously given by

$$\begin{aligned} \cos \Delta\theta_0 &= G_C/y_e = (G_C/M)(M/y_e) \\ &= (M/y_e)/Q. \end{aligned} \tag{7.4}$$

From this we obtain

$$\tan \Delta\theta_0 = \pm (Q^2(y_e/M)^2 - 1)^{\frac{1}{2}}. \tag{7.5}$$

By using (7.3) we obtain

$$(\Delta\omega/\omega_0)_0 = \pm (\tfrac{1}{2}) (y_e/M) (1 - (M/y_eQ)^2)^{\frac{1}{2}} \tag{7.6}$$

or

$$(\Delta\omega/\omega_0)_0 = \pm(\tfrac{1}{2}) (y_e/M) (1 - (G_C/y_e)^2)^{\frac{1}{2}}. \tag{7.7}$$

These equations give the electronic tuning from maximum amplitude of oscillation to zero amplitude of oscillation (extinction).

The equation relating amplitudes may be as easily derived from Fig. 12.

$$G_C^2 + (2M\Delta\omega/\omega)^2 = y_e^2 (2J_1(X)/X)^2 \tag{7.8}$$

at

$$\Delta\omega = 0 \text{ let } X = X_0. \quad \text{Then}$$

$$\Delta\omega/\omega_0 = (y_e/2M) ((2J_1(X)/X)^2 - (2J_1(X_0)/X_0)^2)^{\frac{1}{2}}. \tag{7.9}$$

It is of interest to have the value of $\Delta\omega/\omega_0$ at half the power for $\Delta\omega = 0$. At half power, $X = X_0/\sqrt{2}$, so

$$(\Delta\omega/\omega_0)_{\frac{1}{2}} = (y_e/2M)((2J_1(X_0/\sqrt{2})/(X_0/\sqrt{2}))^2 - (2J_1(X_0)/X_0)^2)^{\frac{1}{2}}. \quad (7.10)$$

For given values of modulation coefficient and V_0, X is a function of the *r-f* gap voltage V and also of drift angle and hence of $\Delta\theta$, or repeller voltage (see Appendix IV). For the fairly large values of θ typical of most reflex oscillators, we can neglect the change in X due directly to changes in $\Delta\theta$, and consider X as a direct measure of the *r-f* gap voltage V. Likewise y_e is a function of drift time whose variation with $\Delta\theta$ can and will be disregarded. Hence from (7.9) we can plot $(X/X_0)^2$ vs. $\Delta\omega/\omega_0$ and regard this as a representation of normalized power vs. frequency.

Let us consider now what (7.3) and (7.9) mean in connection with a given reflex oscillator. Suppose we change the load. This will change Q in (7.3) and X_0 in (7.9). From the relationship previously obtained for the condition for maximum power output, $G_R/y_e = J_0(X_0)$, we can find the value of X_0 that is, X at $\Delta\omega = 0$, for various ratios of G_R/y_e. For $G_R = 0$ (zero resonator loss) the optimum power value of X_0 is 2.4. When there is some resonator loss, the optimum total conductance for best power output is greater and hence the optimum value of X_0 is lower.

In Fig. 13 use is made of (7.3) $\Delta\omega/\omega_0$ in plotted vs. $\Delta\theta$ (which decreases as the repeller is made more negative) for several values of Q, and in Fig. 14, (7.9), is used to plot $(X/X_0)^2$ vs. $(2M/y_e)\Delta\omega/\omega_0$, which is a generalized electronic tuning variable, for several values of X_0. These curves illustrate typical behavior of frequency vs. drift angle or repeller voltage and power vs. frequency for a given reflex oscillator for various loads. In practice, the S shape of the frequency vs. repeller voltage curves for light loads (high Q) is particularly noticeable. The sharpening of the amplitude vs. frequency curves for light loads is also noticeable, though of course the cusp-like appearance for zero load and resonator loss cannot be reproduced experimentally. It is important to notice that while the plot of output vs. frequency for zero load is sharp topped, the plot of output vs. repeller voltage for zero load is not.

Having considered the general shape of frequency vs. repeller voltage curves and power vs. frequency curves, it is interesting to consider curves of electronic tuning to extinction $((\Delta\omega/\omega_0)_0)$ and electronic tuning to half power $((\Delta\omega/\omega_0)_{\frac{1}{2}})$ vs. the loading parameter, $(M/y_eQ) = G_C/y_e$. Such curves are shown in Fig. 15. These curves can be obtained using (7.7) and (7.10). In using (7.10) X can be related to G_C/y_e by the relation previously derived from $2J_1(X)/X = G_C/y_e$ and given in Fig. 5 as a function of X. It is to be noted that the tuning to the half power point, $(\Delta\omega/\omega_0)_{\frac{1}{2}}$, and the tuning to the extinction point, $(\Delta\omega/\omega_0)_0$, vary quite differently with loading.

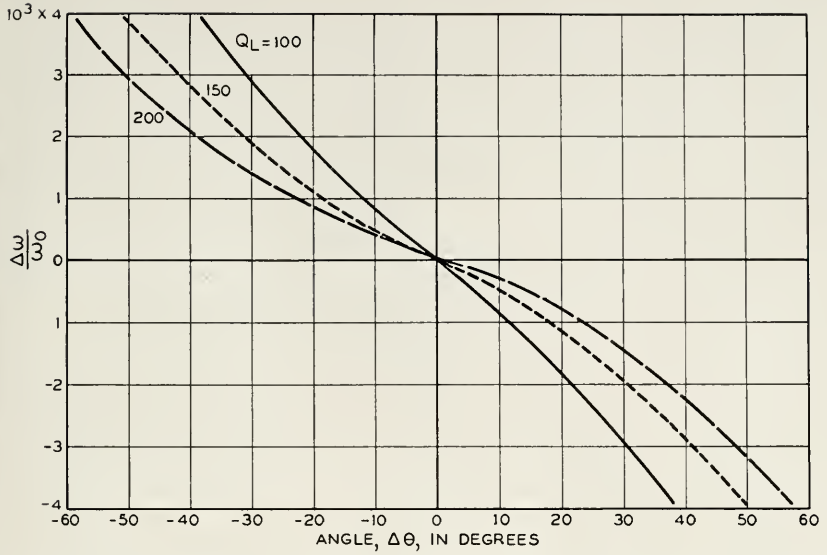


Fig. 13.—A parameter proportional to electronic tuning plotted vs deviation from optimum drift angle $\Delta\theta$ for various values of loaded Q. For lower values of Q, the frequency varies rapidly and almost linearly with $\Delta\theta$. For high values of Q, the frequency curve is S shaped and frequency varies slowly with $\Delta\theta$ for small values of $\Delta\theta$.

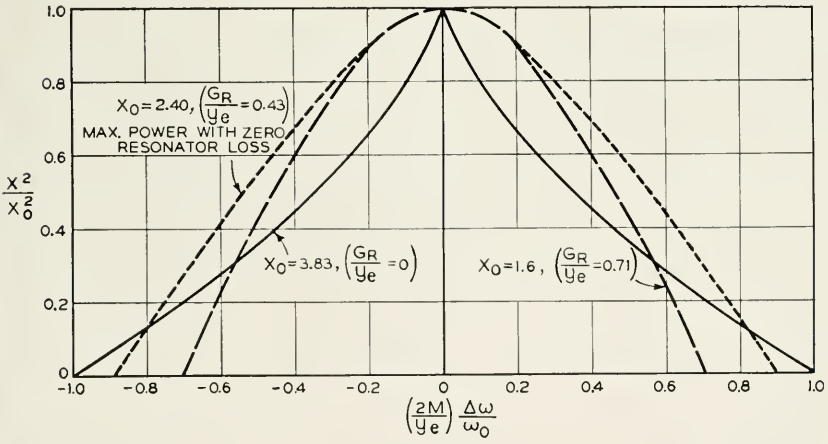


Fig. 14.—The relative power output vs a parameter proportional to the frequency deviation caused by electronic tuning, for various values of load. For zero loss and zero load, the curve is peaked. For zero loss and optimum load, the curve has its greatest width between half power points. For zero loss and greater than optimum load, the curve is narrow.

The quantity

$$(\Delta\omega/\omega_0)_{\frac{1}{2}}$$

has a maximum value at $G_c/y_e = .433(X = 2.40)$, which is the condition for maximum power output when the resonator loss is zero.

In Fig. 11 we have a plot of G_c/y_e vs. G_R/y_e for optimum loading (that is loading to give maximum power for $\Delta\theta = 0$). This, combined with the

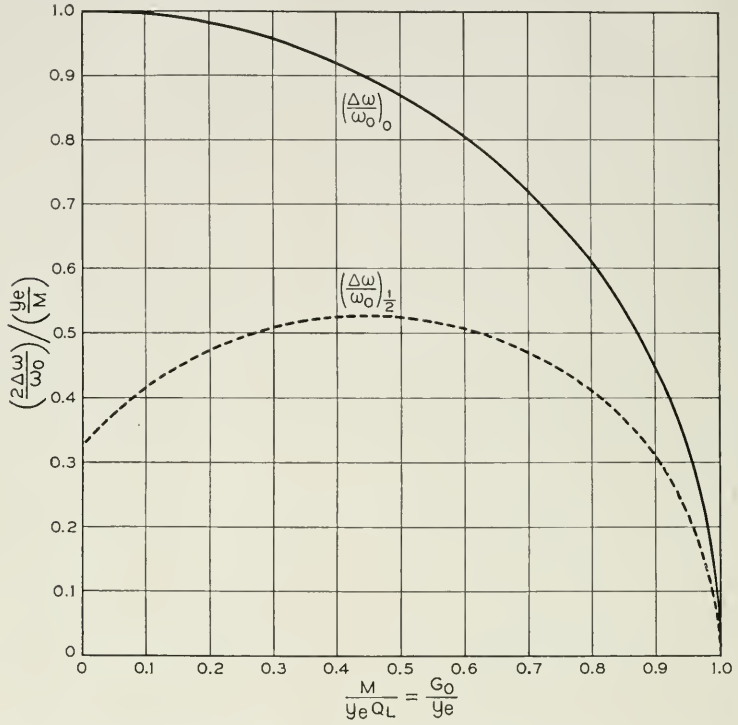


Fig. 15.—A parameter proportional to electronic tuning range vs the ratio of total circuit conductance and small signal electronic admittance. The electronic tuning to extinction $(\Delta\omega/\omega_0)_0$ is more affected by loading than the electronic tuning to half power points $(\Delta\omega/\omega_0)_{\frac{1}{2}}$.

curves of Fig. 15, enables us to draw curves in the case of optimum loading for electronic tuning as a function of the resonator loss. Such curves are shown in Fig. 16.

From Fig. 16 we see that with optimum loading it takes very large resonator losses to affect the electronic tuning range to half power very much, and that the electronic tuning range to extinction is considerably more affected by resonator losses. Turning back to Fig. 7, we see that power is affected even more profoundly by resonator losses. It is interesting to

compare the effect of going from zero loss to a case in which the loss conductance is $\frac{1}{2}$ of the small signal electronic conductance ($G_R = y_e/2$). The table below shows the fraction to which the power or efficiency, the elec-

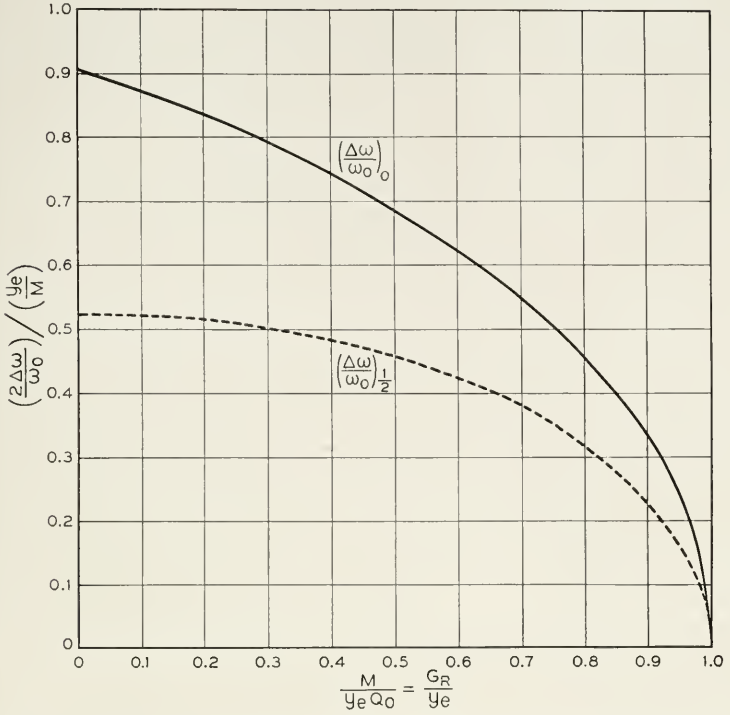


Fig. 16.—The effect of resonator loss on electronic tuning in an oscillator adjusted for optimum power output at the center of the electronic tuning range. A parameter proportional to electronic tuning is plotted vs the ratio of the resonator loss to small signal electronic admittance. The electronic tuning to extinction is more affected than the electronic tuning to half power as the loss is changed.

tronic tuning range to extinction, and the electronic tuning range to half power are reduced by this change.

Power, Efficiency (η)	Electronic Tuning to Extinction $(\Delta\omega/\omega_0)_0$	Electronic Tuning To Half Power $(\Delta\omega/\omega_0)_{\frac{1}{2}}$
.24	.76	.88

From this table it is obvious that efforts to control the electronic tuning by varying the ratio $\frac{G_R}{y_e}$ are of dubious merit.

One other quantity may be of some interest; that is the phase angle of electronic tuning at half power and at extinction. We already have an expression involving $\Delta\theta_0$ (the value at extinction) in (7.4). By taking advantage of (3.10) and (3.8) (Figs. 5 and 11), we can obtain $\Delta\theta_0$ vs. G_C/y_e

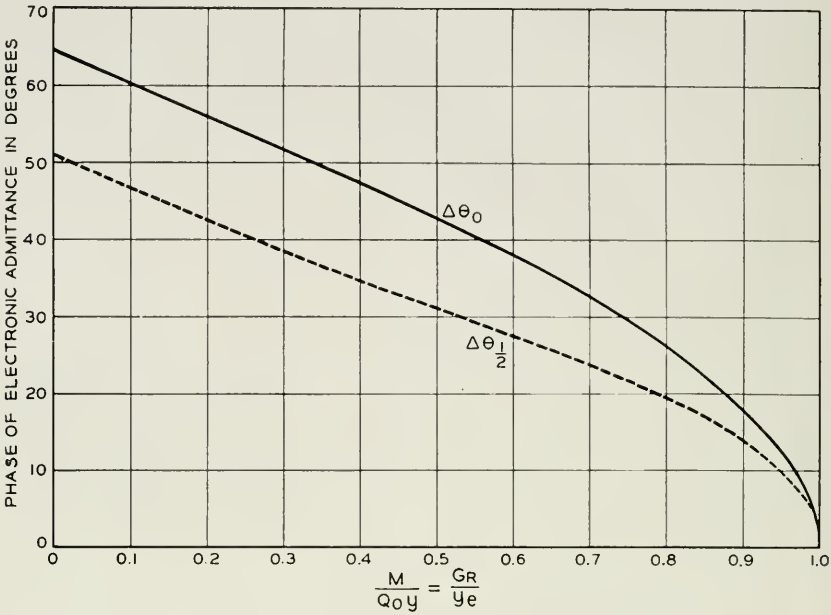


Fig. 17.—The phase of the drift angle for extinction and half power vs the ratio o resonator loss to small signal electronic admittance.

$(=M/Qy_e)$ for optimum loading. By referring to Fig. 12 we can obtain the relation for $\Delta\theta_{\frac{1}{2}}$ (the value at half power)

$$G_C = y_e\,[2J_1(X_0/\sqrt{2})/(X_0/\sqrt{2})]\cos\Delta\theta_{\frac{1}{2}}. \tag{7.11}$$

However, we have at $\Delta\theta = 0$

$$G_C = y_e\,[2J_1(X_0)/X_0]. \tag{7.12}$$

Hence

$$\cos\Delta\theta_{\frac{1}{2}} = \frac{J_1(X_0)}{\sqrt{2}J_1(X_0/\sqrt{2})}. \tag{7.13}$$

Again, from (3.10) and (3.8) we can express X_0 for optimum power at $\Delta\theta = 0$ in terms of G_C/y_e. In Fig. 17, $\Delta\theta_0$ and $\Delta\theta_{\frac{1}{2}}$ are plotted vs. G_C/y_e for optimum loading.

VIII. Hysteresis

All the analysis presented thus far would indicate that if a reflex oscillator is properly coupled to a resistive load the power output and frequency will be single-valued functions of the drift time or of the repeller voltage, as illustrated in Fig. 18. During the course of the development in these laboratories of a reflex oscillator known as the 1349XQ, it was found that even if

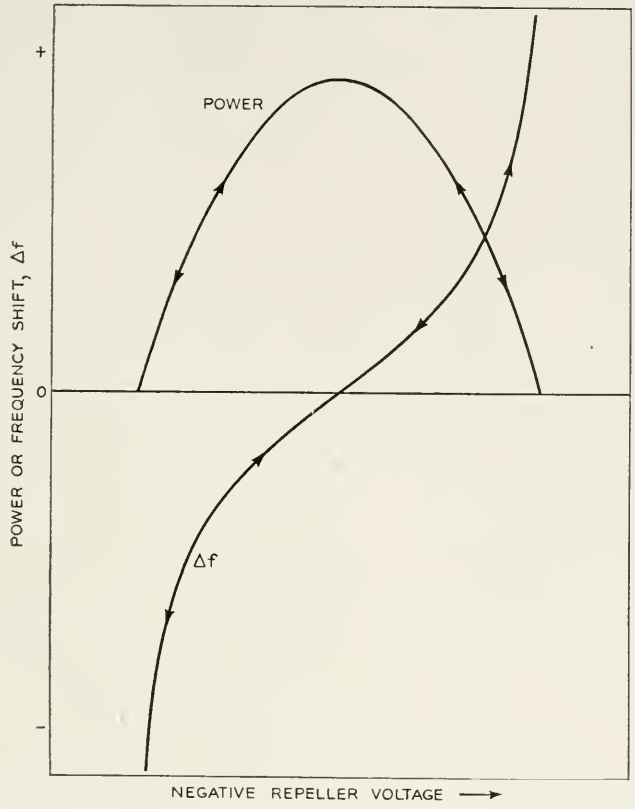


Fig. 18.—Ideal variation of power and frequency with repeller voltage, arbitrary units.

the oscillator were correctly terminated the characteristics departed violently from the ideal, as illustrated in Fig. 19. Further investigation disclosed that this departure was, to a greater or less degree, a general characteristic of all reflex oscillators in which no special steps had been taken to prevent it.

The nature of this departure from expected behavior is that the output is not a single valued function of the repeller voltage, but rather that at a given repeller voltage the output depends upon the direction from which the repel-

ler voltage is made to approach the given voltage. Consider the case illustrated in Fig. 19. The arrows indicate the direction of repeller voltage variation. If we start from the middle of the characteristic and move toward more negative values of repeller voltage, the amplitude of oscillation varies continuously until a critical value is reached, at which a sudden decrease in

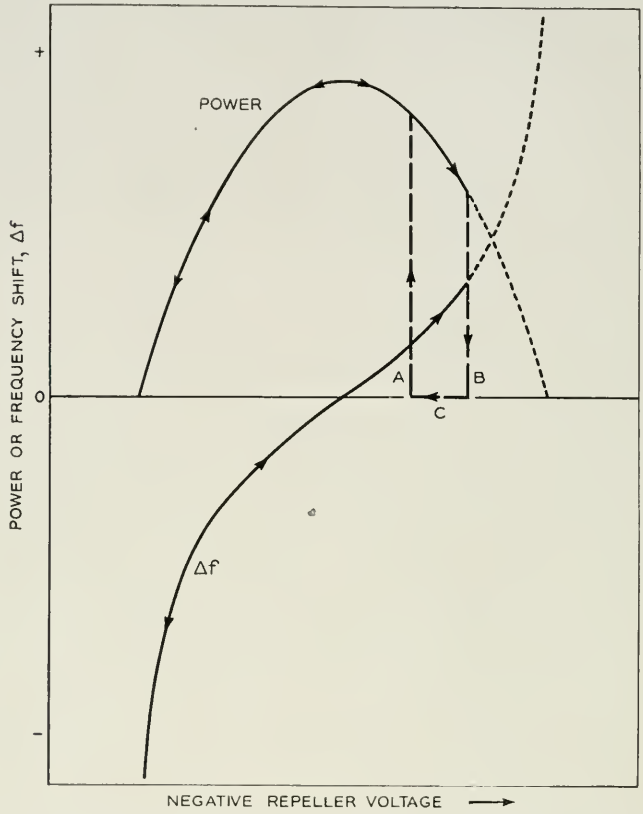


Fig. 19.—A possible variation of power and frequency with repeller voltage when there is electronic hysteresis. The arrows indicate the direction of variation of repeller voltage.

amplitude is observed. This drop may be to zero amplitude as shown or to a finite amplitude. In the latter case the amplitude may again decrease continuously as the repeller voltage is continuously varied to a new critical value, where a second drop occurs, etc. until finally the output falls to zero. In every observed case, even for more than one drop, the oscillation always dropped to zero discontinuously. Upon retracing the repeller voltage variation, oscillation does not restart at the repeller voltage at which it stopped but remains zero until a less negative value is reached, at which point the

oscillation jumps to a large amplitude on the normal curve and then varies uniformly. The discontinuities occur sometimes at one end of the characteristic and sometimes at the other, and infrequently at both. It was first thought that this behavior was caused by an improper load,[7] but further investigation proved that the dependence on the load was secondary and the conclusion was drawn and later verified that the effect had its origin in the electron stream. For this reason the discontinuous behavior was called electronic hysteresis.

In any self-excited oscillator having a simple reasonant circuit, the oscillating circuit may be represented schematically as shown in Fig. 20. Here L and C represent the inductance and capacitance of the oscillator. G_R is a shunt conductance, representing the losses of the circuit, and G_L is the conductance of the load. Henceforth for the sake of convenience we

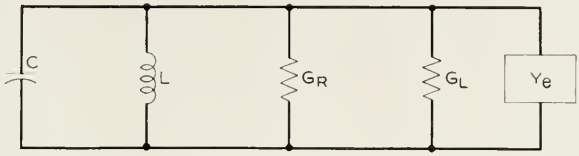


Fig. 20.—Equivalent circuit of reflex oscillator consisting of the capacitance C, inductance L, the resonator loss conductance G_R, the load conductance G_L and the electronic admittance Y_e.

will lump these and call the total G_L. Y_e represents the admittance of the electron stream. Such a circuit has a characteristic transient of the form

$$V = V_0 e^{-\alpha t} e^{j\omega t} \tag{8.1}$$

where

$$\alpha = \frac{G_e + G_L}{2C} \quad \text{and} \quad \omega \doteq \frac{1}{\sqrt{LC}}.$$

Oscillations will build up spontaneously if

$$G_{e0} + G_L < 0. \tag{8.2}$$

For stable oscillation at amplitude V we require

$$G_e[V] + G_L = 0 \tag{8.3}$$

(8.2) and (8.3) state that the amplitude of oscillation will build up until non-linearities in the electronic characteristics reduce the electronic conductance to a value equal and opposite to the total load plus circuit conductance. Thus, in general

$$Y_e = G_{e0}F_1(V) + jB_{e0}F_2(V) \tag{8.4}$$

[7] See Section IX.

where

$$Y_e = G_{e0} + jB_{e0} \tag{8.5}$$

is the admittance for vanishing amplitude, which is taken as a reference value. The foregoing facts are familiar to any one who has worked with oscillators.

Now, condition (8.3) may be satisfied although (8.2) is not. Then an oscillator will not be self-starting, although once started at a sufficiently large amplitude its operation will become stable. An example in common experience is a triode Class C oscillator with fixed grid bias. In such a case

$$F(V_1) > F(0) \tag{8.6}$$

holds for some V_1.

As an example of normal behavior, let us assume that $F(V)$ is a continuous monotonically decreasing function of increasing V, with the reference value of V taken as zero. Then the conductance, $G_e = G_{e0}F(V)$ will vary with V as shown in Fig. 21. Stable oscillation will occur when the amplitude V_1 has built up to a value such that the electronic conductance curve intersects the horizontal line representing the load conductance, G_L. G_{e0} is a function of one or more of the operating parameters such as the electron current in the vacuum tube. If we vary any one of these parameters indicated as X_n the principal effect will be to shrink the vertical ordinates as shown in Fig. 21 and the amplitude of oscillation will assume a series of stable values corresponding to the intercepts of the electronic conductance curves with the load conductance. If, as we have assumed, $F(V)$ is a monotonically decreasing function of V, the amplitude will decrease continuously to zero as we uniformly vary the parameter in such a direction as to decrease G_{e0}. Zero amplitude will, of course, occur when the curve has shrunk to the case where $G_{e0} = G_L$. Under these conditions the power output, $\frac{1}{2}G_L V^2$, will be a single value function of the parameter as shown in Fig. 22 and no hysteresis will occur.

Suppose, however, that $F(V)$ is not a monotonically decreasing function of V but instead has a maximum so that $G_{e0}F(V)$ appears as shown in Fig. 23. In this case, if we start with the condition indicated by the solid line and vary our parameter X in such a direction as to shrink the curve, the amplitude will decrease smoothly until the parameter arrives at a value of X_5 corresponding to amplitude V_5 at which the load line is tangent to the maximum of the conductance curve. Further variation of X in the same direction will cause the amplitude to jump to zero. Upon reversing the direction of the variation of the parameter, oscillation cannot restart until X arrives at a value X_4 such that the zero amplitude conductance is equal to the load conductance. When this occurs the amplitude will suddenly jump to the

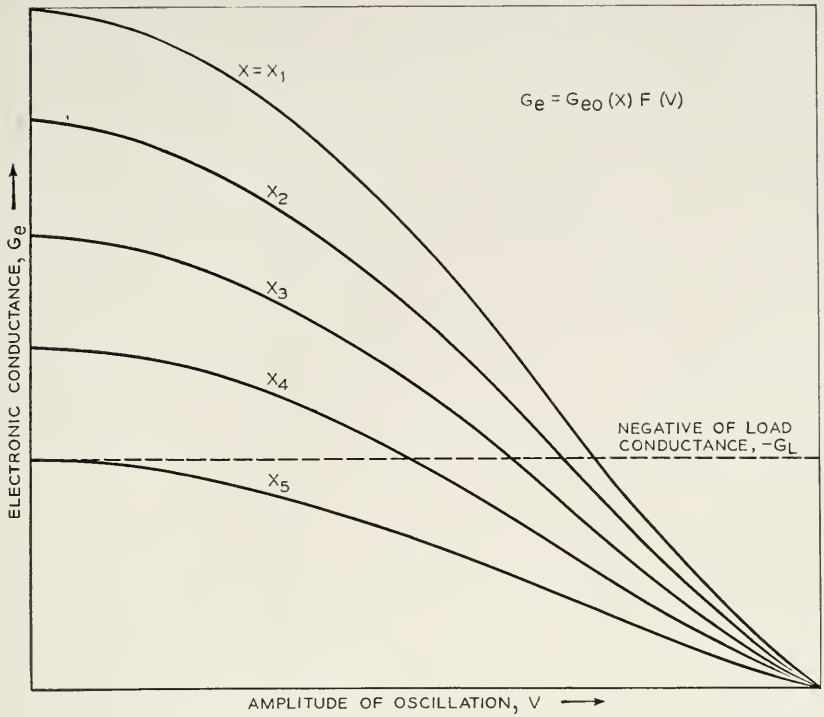


Fig. 21.—A possible variation of electronic conductance with amplitude of oscillation for the general case of an oscillator. Arbitrary units are employed. Different curves correspond to several values of a parameter X which determines the small signal values of the conductance. The load conductance is indicated by the horizontal line. Stable oscillation for any given value of the parameter X occurs at the intersection of the electronic conductance curve with the load line G_L.

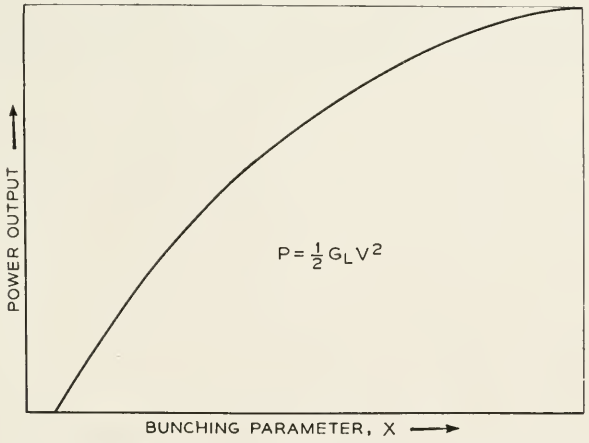


Fig. 22.—Variation of power output with the parameter X when the conditions of Fig. 21 apply.

value V_4. Under these conditions the power output will appear as shown in Fig. 24, in which the hysteresis is apparent.

Let us now consider the conditions obtaining in a reflex oscillator. Fig. 1 shows a schematic diagram of a reflex oscillator. This shows an electron gun which projects a rectilinear electron stream across the gap of a resonator.

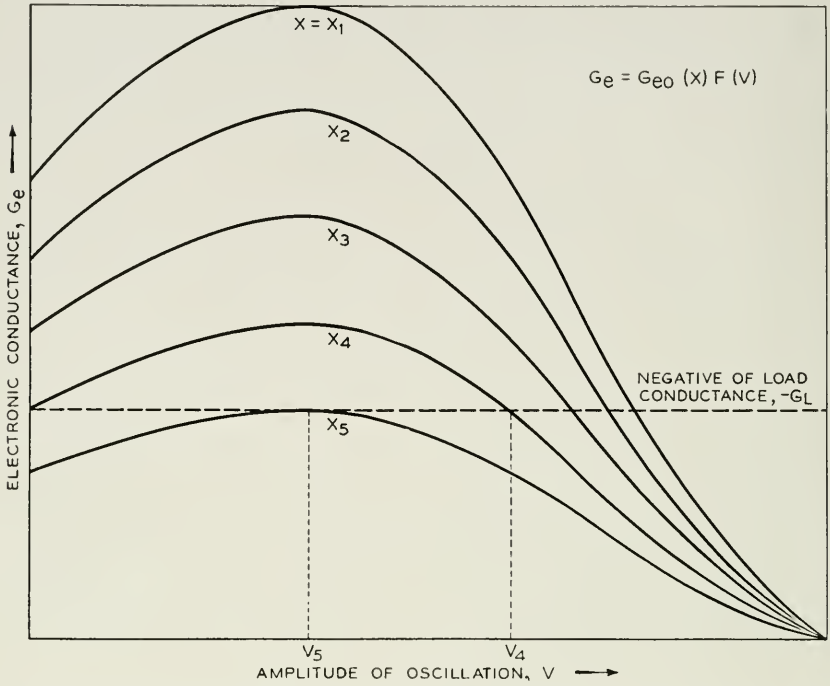


Fig. 23.—Variation of electronic conductance with amplitude of oscillation of a form which will result in hysteresis. The parameter X determines the small signal value of the conductance. The horizontal line indicates the load conductance.

After the beam passes through this gap it is retarded and returned by a uniform electrostatic field. If we carry out an analysis to determine the electronic admittance which will appear across the gap if the electrons make one round trip, we arrive at expression 2.2 which may be written

$$Y_e = \frac{I_0 \beta^2 \theta}{V_0} \frac{J_1(X)}{X} [\sin \theta + j \cos \theta] \tag{8.7}$$

where

$$X = \frac{\theta \beta V}{2 V_0}.$$

This admittance will be a pure conductance if $\theta = \theta_0 = (n + \frac{3}{4})\,2\pi$. As we have seen, in an oscillator designed specifically for electronic tuning, n usually has a value of 3 or greater and the variations $\Delta\theta$ from θ arising from

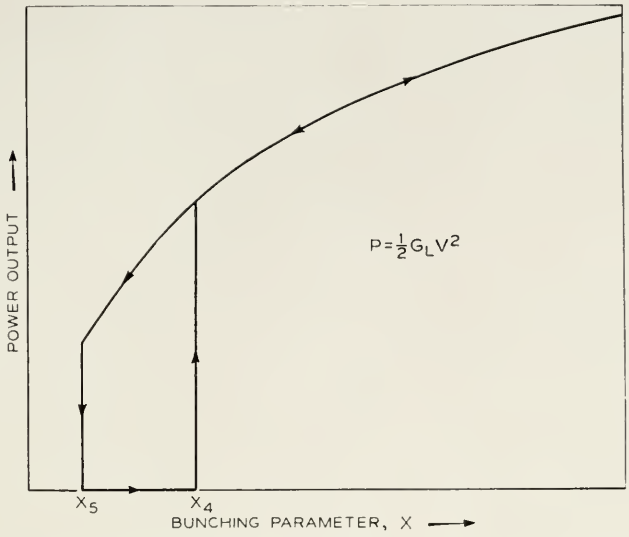


Fig. 24.—A curve of power output vs parameter X resulting from the conductance curves shown in Fig. 23 and illustrating hysteresis.

repeller voltage variation are sufficiently small so that the effect of $\Delta\theta$ in varying X may be neglected. In this case we may write

$$G_e = -y_e \frac{2J_1(CV)}{CV} \cos \Delta\theta$$

$$y_e = \frac{i_0 \beta^2 \theta_0}{2V_0} \tag{8.8}$$

$$C = \frac{\theta_0 \beta}{2V_0}.$$

The parameter which we vary in obtaining the repeller characteristic of the tube is $\Delta\theta$. The variation of this parameter is produced by shifting the repeller voltage V_R from the value V_{R0} corresponding to the transit angle θ_0. Since as is shown, Fig. 25a, $\frac{J_1(CV)}{CV}$ decreases monotonically as V increases, no explanation of hysteresis is to be found in this expression. Fig. 25b shows the smooth symmetrical variation of output with repeller voltage about the value for which $\Delta\theta = 0$ which is to be expected.

Now suppose a second source of conductance G_{e2} exists whose amplitude function is of the form illustrated in Fig. 26a. Let us suppose that for the

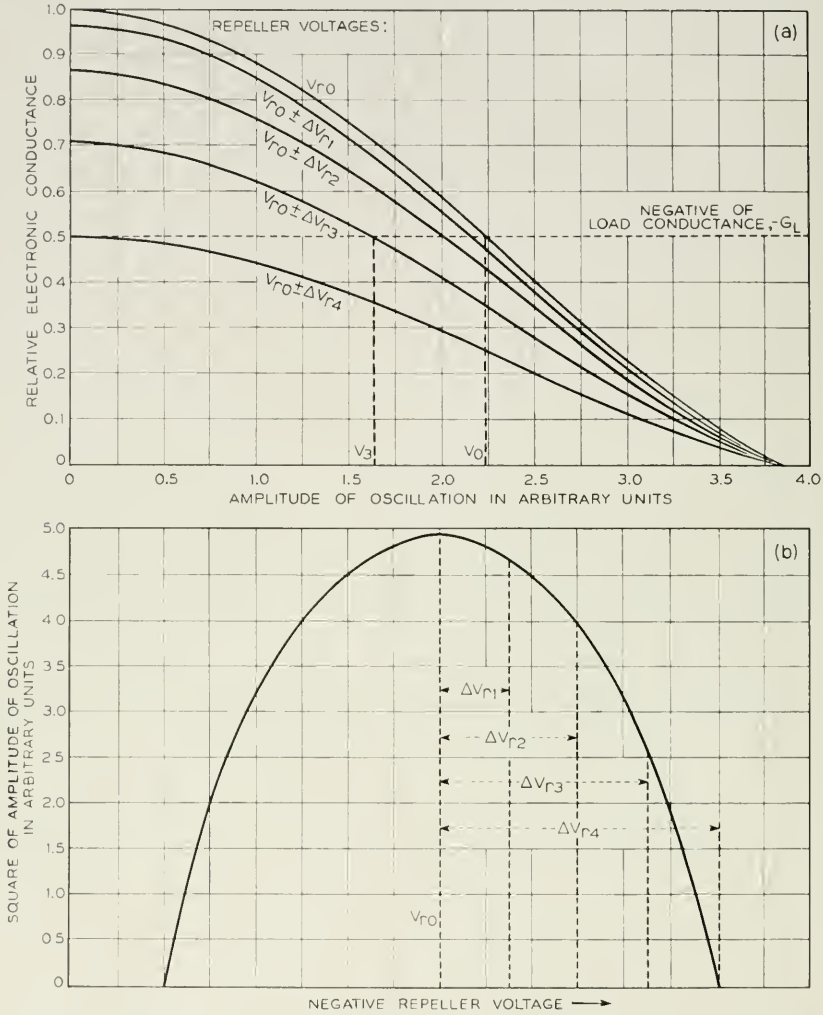


Fig. 25.—*a*. Variation of electronic conductance with amplitude of oscillation for an ideal oscillator. The parameter controlling the small signal electronic conductance is the repeller voltage which determines the transit angle in the repeller region. The horizontal line indicates the load conductance.

b. The variation of power output with the repeller voltage which results from the characteristics of Fig. 25a.

value of θ_0 assumed the phase of this conductance is such as to oppose G_{e1}, G_{e2} may or may not be a function of $\Delta\theta$. For the sake of simplicity let us assume that G_{e2} varies with $\Delta\theta$ in the same way as G_{e1}. The total conduc-

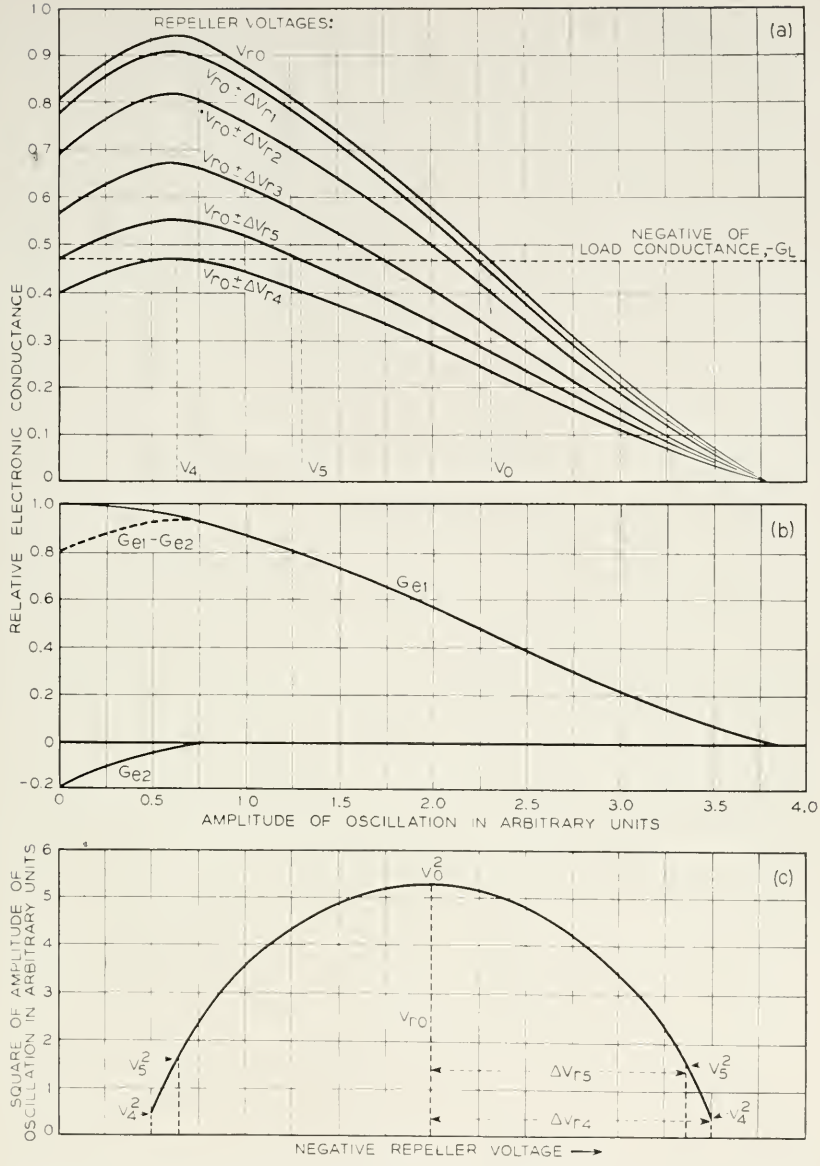


Fig. 26.—*a*. Curve G_{e1} shows the variation of electronic conductance with amplitude of oscillation for an ideal reflex oscillator. Curve G_{e2} represents the variation of a second source of electronic conductance with amplitude. The difference of these two curves indicated G_{e1}-G_{e2} shows the variation of the sum of these two conductance terms with amplitude.

b. Electronic conductance vs amplitude of oscillation when two conductance terms exist whose variation with repeller voltage is the same.

c. Power output vs repeller voltage for a reflex oscillator in which two sources of conductance occur varying with amplitude as shown in Fig. 26b.

tance $G_e = G_{e1} - G_{e2}$ will appear as shown. As the repeller voltage is varied from the optimum value the conductance curve will shrink in proportion to $\cos \Delta\theta$, and the amplitude of oscillation for each value of $\Delta\theta$ will adjust itself to the value corresponding to the intersection of the load line and the conductance plot as shown in Fig. 26b. When the load line becomes tangent, as for amplitude V_4, further variation of the repeller voltage in the same direction will cause oscillation to jump from V_4 to zero amplitude. Correspondingly, on starting oscillation will restart with a jump to V_3. Hence, two sources of conductance varying in this way will produce conditions previously described, which would cause hysteresis as shown in Fig. 26c. The above assumptions lead to hysteresis symmetrically disposed about the optimum repeller voltage. Actually, this is rarely the case, but the explanation for this will be deferred.

Fig. 27 shows repeller characteristics for an early model of a reflex oscillator designed at the Bell Telephone Laboratories. The construction of this oscillator was essentially that of the idealized oscillator of Fig. 1 upon which the simple theory is based. However, the repeller characteristics of this oscillator depart drastically from the ideal. It will be observed that a double jump occurs in the amplitude of oscillation. The arrows indicate the direction of variation of the repeller voltage. The variation in the frequency of oscillation is shown, and it will be observed that this also is discontinuous and presents a striking feature in that the rate of change of frequency with voltage actually reverses its sign for a portion of the range. A third curve is shown which gives the calculated phase $\Delta\theta$ of the admittance arising from drift in the repeller field. This lends very strong support to the hypothesis of the existence of a second source of conductance, since this phase varies by more than 180°, so that for some part of the range the repeller conductance must actually oppose oscillation. The zero value phase is arbitrary, since there is no way of determining when the total angle is $(n + \frac{3}{4})2\pi$.

Having recognized the circumstances which can lead to hysteresis in the reflex oscillator, the problem resolves itself into locating the second source of conductance and eliminating it.

A number of possible sources of a second conductance term were investigated in the particular case of the 1349 oscillator, and most were found to be of negligible importance. It was found that at least one important second source of conductance arose from multiple transits of the gap made by electrons returning to the cathode region. In the case of the 1349 a design of the electron optical system which insured that the electron stream made only one outgoing and one return transit of the gap eliminated the hysteresis in accordance with the hypothesis.

Inasmuch as multiple transits appear to be the most common cause for hysteresis in reflex oscillator design, it seems worthwhile to obtain a more detailed understanding of the mechanism in this case. Other possible

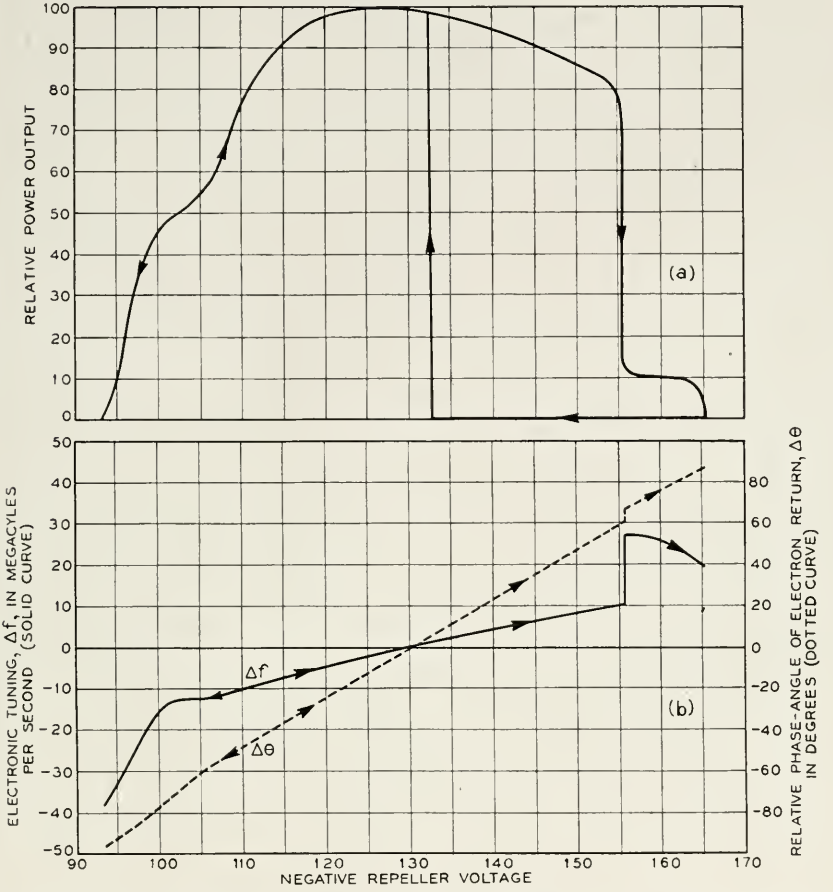


Fig. 27.—Amplitude, frequency and transit phase variation with the repeller voltage obtained experimentally for a reflex oscillator exhibiting electronic hysteresis. The arrows indicate the direction of variation of the repeller voltage.

mechanisms such as velocity sorting on the repeller will give rise to similar effects and can be understood from what follows.

In the first order theory, the electrons which have retraversed the gap are conveniently assumed to vanish. Actually, of course, the returning stream is remodulated and enters the cathode space. Unfortunately, the

conditions in the cathode region are very complex, and an exact analysis would entail an unwarranted amount of effort. However, from an approximate analysis one can obtain a very simple and adequate understanding of the processes involved.

Let us examine the conditions existing after the electrons have returned through the gap of the idealized reflex oscillator. In the absence of oscillation, with an ideal rectilinear stream and ideally fine grids all the electrons which leave the cathode will return to it. When oscillation exists all electrons which experience a net gain of energy on the two transits will be captured by the cathode, while those experiencing a net loss will not reach it, but instead will return through the gap for a third transit, etc. In a practical oscillator even in the absence of oscillation only a fraction of the electrons which leave the cathode will be able to return to the cathode, because of losses in axial velocity produced by deflections by the grid wires and various other causes. As a result, it will not be until an appreciable amplitude of oscillation has been reached that a major proportion of the electrons which have gained energy will be captured by the cathode. On the other hand, there will be an amplitude of oscillation above which no appreciable change in the number captured will occur.

The sorting action which occurs on the cathode will produce a source of electronic admittance. Another contribution may arise from space charge interaction of the returning bunched beam with the outgoing stream. A third component arises from the continued bunching resulting from the first transit of the gap. From the standpoint of this third component the reflex oscillator with multiple transits suggests the action of a cascade amplifier. The situation is greatly complicated by the nature of the drift field in the cathode space. All three mechanisms suggested above may combine to give a resultant second source. Here we will consider only the third component. Consider qualitatively what happens in the bunching action of a reflex oscillator. Over one cycle of the r.f. field, the electrons tend to bunch about the electron which on its first transit crosses the gap when the field is changing from an accelerating to a decelerating value. The group recrosses the gap in such a phase that the field extracts at least as much energy from every electron as it gave up to any electron in the group. When we consider in addition various radial deflections, we see that very few of the electrons constituting this bunch can be lost on the cathode.

Although it is an oversimplification, let us assume that we have a linear retarding field in the cathode region and also that none of the electrons are intercepted on the cathode. To this order of approximation a modified cascade bunching theory would hardly be warranted and we will consider only that the initial bunching action is continued. Under these conditions,

we can show that the admittance arising on the third transit of the gap will have the form

$$Y_e' = +I_0' \frac{\beta^2 \theta_t}{V_0} \frac{J_1\left(\frac{\theta_t \beta V}{2V_0}\right)}{\frac{\theta_t \beta V}{2V_0}} [\sin \theta_t + j \cos \theta_t] \tag{8.9}$$

where I_0' is the effective d.c. contributing to the third transit. $\theta_t = \theta + \theta_c$ is the total transit angle made up of the drift angle in the repeller space, θ, and the drift angle in the cathode space θ_c. As before, assume that the small changes in θ_t caused by the changing repeller voltage over the electronic tuning range exercise an appreciable effect only in changing the sine and cosine terms. Then we may write

$$Y_e' = G_e' + jB_e' = y_e' \frac{2J_1(C_2 V)}{C_2 V} [\sin \theta_t + j \cos \theta_t] \tag{8.10}$$

where

$$C_2 = \frac{\theta_{t0} \beta V}{2V_0}, \quad \text{and} \quad y_e' = \frac{\beta^2 I_0' \theta_{to}}{2V_0}.$$

If $\Delta\theta = \theta_t - \theta_{t0}$

$$G_e' = y_e' \frac{2J_1(C_2 V)}{C_2 V} [\sin \theta_{t0} \cos \Delta\theta_t + \cos \theta_{t0} \sin \Delta\theta_t]. \tag{8.11}$$

Now

$$\begin{aligned} \Delta\theta &= \omega_0 \tau \frac{\Delta V_R}{V_R + V_0} + \Delta\omega\tau_0 \\ \Delta\theta_c &= \Delta\omega\tau_c \\ \Delta\theta_t &= \omega_0 \tau \frac{\Delta V_R}{V_R + V_0} + \Delta\omega\tau_0 + \Delta\omega\tau_c . \end{aligned} \tag{8.12}$$

We observe that the phase angle of the admittance arising on the third transit varies more rapidly with repeller voltage (i.e., frequency) than the phase angle of the second transit admittance. This is of considerable importance in understanding some of the features of hysteresis.

Let us consider (8.11) for some particular values of θ_c or θ_t. We remember that θ_t is greater than θ and hence $C_2 > C_1$. Since this is so, the limiting function $\frac{2J_2(C_2 V)}{C_2 V}$ will become zero at a lower value of V than $\frac{2J_1(C_1 V)}{C_1 V}$. We will consider two cases $\theta_t = (n + \frac{1}{4})2\pi$ and $\theta_t = (n + \frac{3}{4})2\pi$. These

correspond respectively to a conductance aiding and bucking the conductance arising on the first return. In case 1 we have

$$G_e' = -y_e' \frac{2J_1(C_2 V)}{C_2 V} \cos \Delta\theta_t \tag{8.13}$$

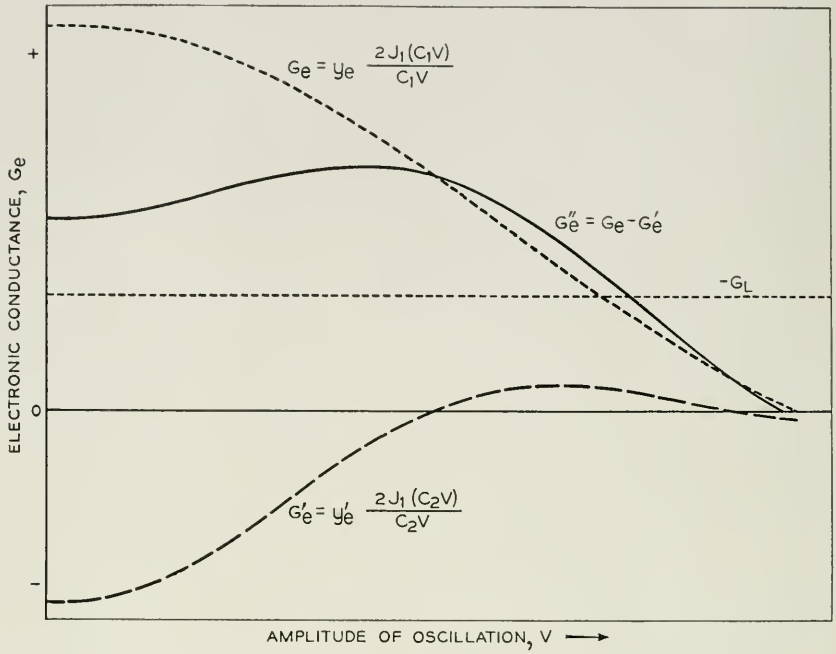


Fig. 28.—Theoretically derived variation of electronic conductance with amplitude of oscillation. Curve G_e represents conductance arising from drift action in the repeller space. Curve G_e' represents the conductance arising from continuing drift in the cathode region. G_e'' represents the conductance variation with amplitude which will result if G_e and G_e are in phase opposition.

and case 2

$$G_e' = +y_e' \frac{2J_1(C_2 V)}{C_2 V} \cos \Delta\theta_t \,. \tag{8.14}$$

Figure 28 illustrates case (2) and Fig. 29 case (1). If $\cos \Delta\theta_t$ and $\cos \Delta\theta$ varied in the same way with repeller voltage, the resultant limiting function would shrink without change in form as the repeller voltage was varied, and it is apparent that Fig. 28 would then yield the conditions for hysteresis and Fig. 29 would result in conditions for a continuous characteristic. If Fig. 28 applied we should expect hysteresis symmetrical about the optimum repeller voltage. We recall, however, that in Fig. 27 hysteresis

occurred only on one end of the repeller characteristic and was absent on the other. The key to this situation lies in the fact that $\Delta\theta_t$ and $\Delta\theta$ do not vary in the same way when the repeller voltage is changed and the frequency shifts as shown in (8.12). As a result, the resulting limiting function does not shrink uniformly with repeller voltage, since the contribution G_e' changes more rapidly than G_e. Hence we should need a continuous series of pictures of the limiting function in order to understand the situation completely.

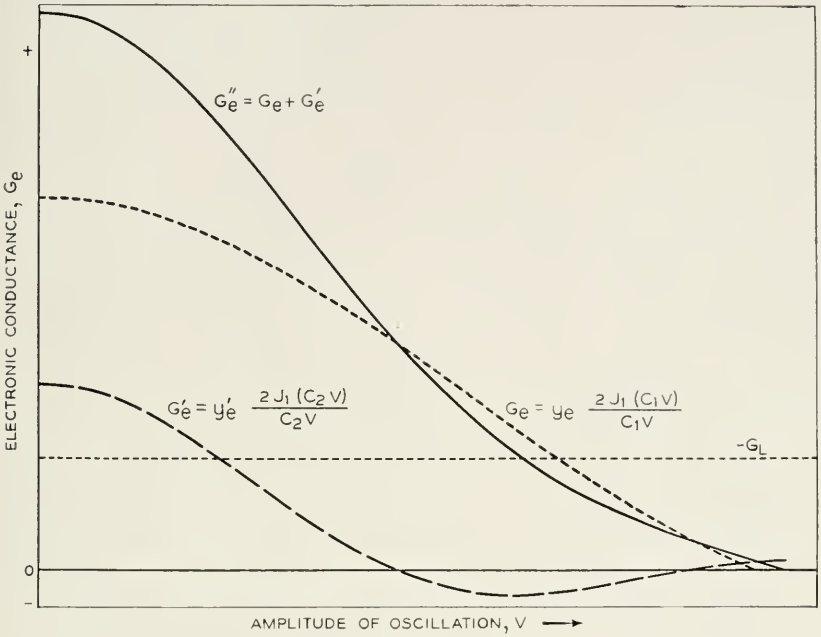


Fig. 29.—Theoretically derived curves of electronic conductance vs amplitude of oscillation. Curve G_e'' shows the variation of the resultant electronic conductance when the repeller space contribution and the cathode space contribution are in phase addition.

Suppose we consider Fig. 29 and again assume in the interests of simplicity that $\Delta\theta_t$ and $\Delta\theta$ vary at the same rate. In this case we observe that in the region aa' the conductance varies very rapidly with amplitude. This would imply that in this region the output would tend to be independent of the repeller voltage. If we refer again to Fig. 27 we observe that the output is indeed nearly independent of the repeller voltage over a range.

We see that these facts all fit into a picture in which, because of the more rapid phase variation of θ_t than θ with repeller voltage, the limiting function at one end of the repeller voltage characteristic has the form of Fig. 28. accounting for the hysteresis, and at the other end has the form of Fig. 29,

accounting for the relative independence of the output on the repeller voltage.

In what has been given so far we have arrived qualitatively at an explanation for the variation of the amplitude. There remains the explanation for the behavior of the frequency. In this case we plot susceptance as a function of amplitude and, as in the case of the conductance, there will be several contributions. The primary electronic susceptance will be given by

$$B_e = y_e \frac{2J_1(CV)}{CV} \sin \theta. \tag{8.15}$$

Hence, as we vary the parameter $\Delta\theta$ by changing the repeller voltage the susceptance curve swells as the conductance curve shrinks. The circuit condition for stable oscillation is that

$$B_e + 2j\Delta\omega C = 0. \tag{8.16}$$

A second source of susceptance will arise from the continuing drift in the cathode space. Referring to equation (8.10) we see that this will have the form

$$B_e' = y_e' \frac{2J_1(C_2 V)}{C_2 V} \cos \theta_t \tag{8.17}$$

and corresponding to equation (8.11) we write

$$B_e' = y_e' \frac{2J_1(C_2 V)}{C_2 V} [\cos \theta_{t0} \cos \Delta\theta_t - \sin \theta_{t0} \sin \Delta\theta_t]. \tag{8.18}$$

Consider the functions given by (8.18) for values of $\theta_t = (n + \frac{1}{4})2\pi$ and $(n + \frac{3}{4})2\pi$ as functions of V. These are the extreme values which we considered in the case of the conductance. The ordinates of these curves give the frequency shift as a function of the amplitude.
In case 1 we have

$$B_e' = -y_e' \frac{2J_1(C_2 V)}{C_2 V} \sin \Delta\theta_t \tag{8.19}$$

and case 2

$$B_e' = y_e' \frac{2J_1(C_2 V)}{C_2 V} \sin \Delta\theta_t \,. \tag{8.20}$$

The total susceptance will be the sum of the susceptance appearing across the gap as a result of the drift in the repeller space and the susceptance which appears across the gap as a result of the cascaded drift action in the repeller region and the cathode region. If $\sin \Delta\theta_t$ and $\sin \Delta\theta$ varied in the same way with the repeller voltage, the total susceptance would expand

or contract without change in form as the repeller voltage was varied. In Figs. 30 and 31 a family of susceptance curves are shown corresponding respectively to cases 1 and 2 above for various values of $\Delta\theta_t$, assuming that $\Delta\theta_t$ and $\Delta\theta$ vary in the same way with the repeller voltage. As the

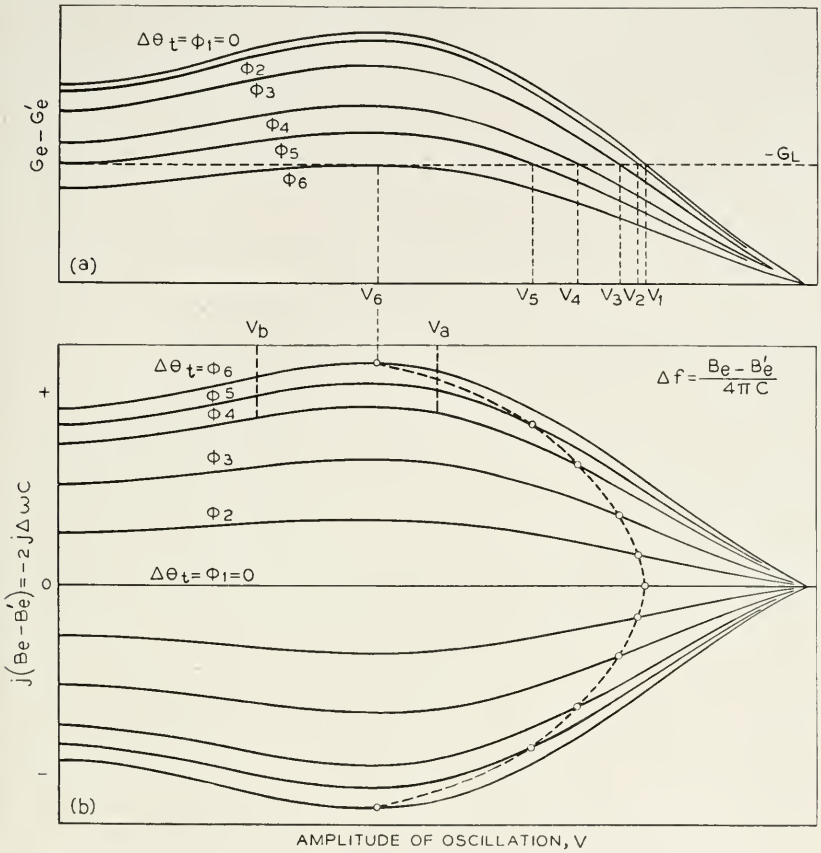


Fig. 30.—*a*. Theoretical variation of electronic conductance vs amplitude of oscillation in the case in which two components are in phase opposition. The parameter is the repeller transit phase. It is assumed that the two contributions have the same variation with this phase.

b. Susceptance component of electronic admittance as a function of amplitude for the case of phase opposition given in Fig. 30a. The parameter is the repeller phase. The dashed line shows the variation of amplitude with the susceptance shift.

repeller voltage is varied the amplitude of oscillation will be determined by the conductance limiting function. In the case of the susceptance we cannot determine the frequency from the intersection of the curve with a load line. The frequency of oscillation will be determined by the drift angle and the amplitude of oscillation. The amplitude variation with

angle may be obtained from Fig. 30a, which gives the conductance family. This gives the frequency variation with angle indicated by the curve con-

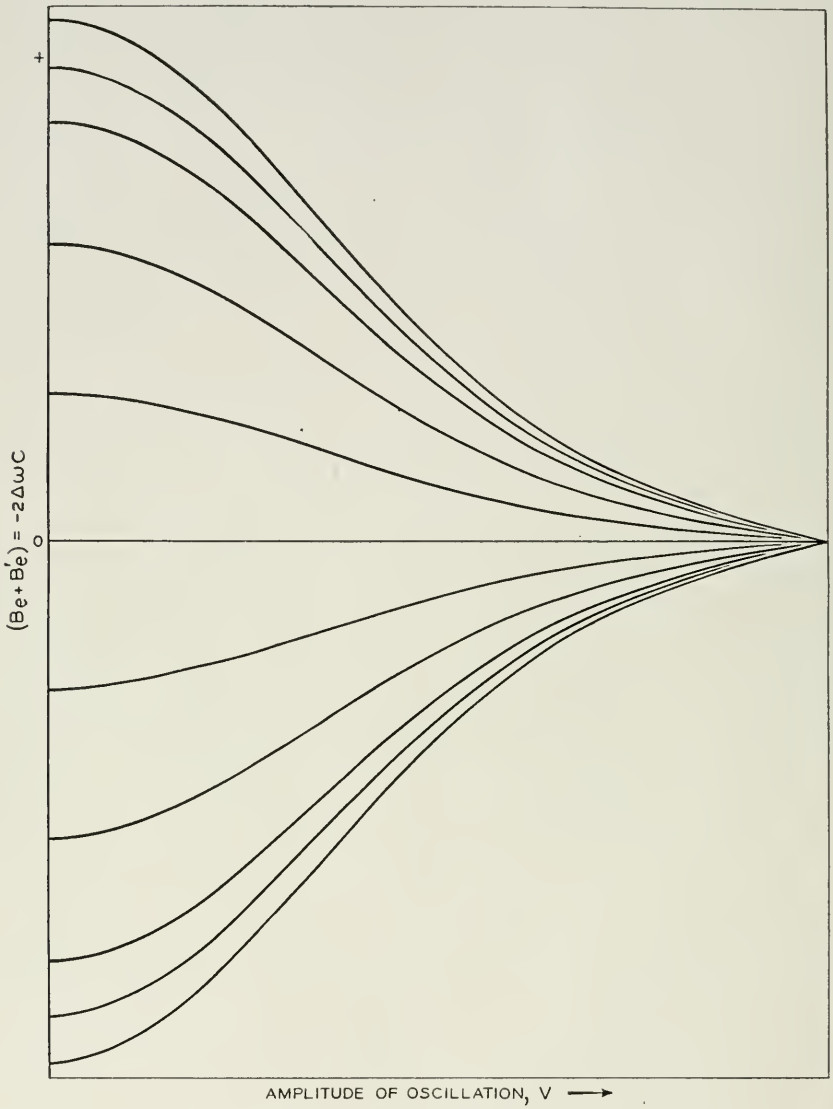


Fig. 31.—Theoretical variation of the susceptance components of electronic admittance vs amplitude of oscillation for the case in which two components of electronic susceptance are in phase addition.

necting the dots of Fig. 30b. On the assumption that $\Delta\theta_t$ and $\Delta\theta$ vary at the same rate with repeller voltage a symmetrical variation about $\Delta\theta = 0$ will occur as shown in Fig. 30b. However, from the arguments used con-

cerning the conductance the actual case would involve a transition from the situation of Fig. 30b to that of Fig. 31. If a discontinuity in amplitude occurs in which the amplitude does not go to zero, it will be accompanied by a discontinuity in frequency, since the discontinuity in amplitude in general will cause a discontinuity in the susceptance. If this discontinuity in susceptance occurs between values of the amplitude such as V_a and V_b of Fig. 30, we observe that the direction of the frequency jump may be opposite to the previous variation. We also observe that if the rate of change of susceptance with amplitude is greater than the rate of change of susceptance with $\Delta\theta$, then in regions such as that lying between zero amplitude of V_b the rate of change of frequency with $\Delta\theta$ may reverse its direction.

One can see that because of the longer drift time contributing to the third transit the conductance arising on the third transit may be of the same order as that arising on the second transit. In oscillators in which several repeller modes, i.e., various numbers of drift angles, may be displayed, one finds that the hysteresis is most serious for the modes with the fewest cycles of drift in the repeller space. One might expect this, since for these modes the contribution from the cathode space is relatively more important.

Some final general remarks will be made concerning hysteresis. One thing is obvious from what has been said. With the admittance conditions as depicted, if all the electronic operating conditions are fixed and the load is varied hysteresis with load can exist. This was found to be true experimentally, and in the case of oscillators working into misterminated long lines it can produce disastrous effects. Where hysteresis is severe enough, it will be found that what we have chosen to call the sink margin will be much less than the theoretically expected value. An illustration of this is given in Fig. 109.

The explanation which we have given for the hysteresis in the reflex oscillator depends upon the existence of two sources of conductance. This was apparently a correct assumption in the case studied, since the elimination of the second source also eliminated the hysteresis. It is possible, however, to obtain hysteresis in a reflex oscillator with only a single source. This can occur if the phase of the electronic admittance is not independent of the amplitude. Normally, in adjusting the repeller voltage the value is chosen for the condition of maximum output. This means that the drift angle is set to a value to give maximum conductance for large amplitude. If the drift angle is then a function of the amplitude, this will mean that for small amplitude it will no longer be optimum. Thus, although the limiting function $\frac{J_1(CV)}{CV}$ tends to increase the electronic conductance as the amplitude declines, the phase factor will oppose this increase. If the phase factor depended sufficiently strongly on the amplitude, the decrease in G_e caused by

the phase might outweigh the increase due to the function $\frac{J_1(C_1V)}{C_1V}$. As a result the conductance might have a maximum value for an amplitude greater than zero, leading to the conditions shown in Fig. 23, under which hysteresis can exist.

The first order theory for the reflex oscillator does not predict such an effect, since the phase is independent of amplitude. The second order theory gives the admittance as

$$Y_e = \frac{\beta_2 I_0 \theta}{2V_0} \cdot \frac{2J_1(X)}{X} e^{-j(\theta-(\pi/2))} \left(1 - \frac{1}{4\theta^2} \cdot \left[\tfrac{1}{2}X^2(X^2+1) - X^3\frac{J_0(X)}{X} - \frac{j}{2\theta}(2-X^2) - X\frac{J_0(X)}{J_1(X)}\right]\right) \tag{8.21}$$

The quantity appearing outside the brackets is the admittance given by the first order theory. The second order correction contains real and imaginary parts which are functions of X and hence of the amplitude of oscillation. Thus, for fixed d-c conditions the admittance phase depends upon the amplitude of oscillation and hence hysteresis might occur. It should be observed that the correction terms are important only for small values of the transit angle θ. In particular, this explanation would not suffice for the case described earlier since the design employed which eliminated the hysteresis left the variables of equation (8.21) unchanged.

IX. Effect of Load

So far we have considered the reflex oscillator chiefly from the point of view of optimum performance; that is, we have attempted chiefly to evaluate its performance when it is used most advantageously. There has been some discussion of non-optimum loading, but this has been incidental to the general purpose of the work. Oscillators frequently are worked into other than optimum loads, sometimes as a result of incorrect adjustment, sometimes through mistakes in design of equipment and quite frequently by intention in order to take advantage of particular properties of the reflex oscillator when worked into specific non-optimum loads.

In this section we will consider the effects of other than optimum loads on the performance of the reflex oscillator. We may divide this discussion into two major subdivisions classified according to the type of load. The first type we call fixed element loads, and the second variable element loads. The first type is constructed of arbitrary passive elements whose constants are independent of frequency. The second category includes loads constructed of the same type of elements but connected to the oscillator by lines of sufficient length so that the frequency variation of the load admittance is appreciably modified by the line.

A. *Fixed Element Loads*

In this discussion it will be assumed initially that $\Delta\theta$, the phase angle of $-Y_e$, is not affected by frequency. The results will be extended later to account for the variation of $\Delta\theta$ with frequency. A further simplification is the use of the equivalent circuit of Fig. 118, Appendix I. Initially, the output circuit loss, R, will be taken as zero, so the admittance at the gap will be

$$Y_c = G_R + 2jM\Delta\omega/\omega + Y_L/N^2. \tag{9.1}$$[8]

Here, G_R is the resonator loss conductance, M is the resonator characteristic admittance, and Y_L is the load admittance.

We will now simplify this further by letting $G_R = 0$

$$Y_c = 2jM\Delta\omega/\omega + Y_L/N^2. \tag{9.2}$$

From Fig. 12 we see

$$G_L/N^2 = y_e[2J_1(X)/X] \cos \Delta\theta \tag{9.3}$$

$$\frac{B_L}{N^2} + \frac{2M\Delta\omega}{\omega} = -y_e[2J_1(X)/X] \sin \Delta\theta\,. \tag{9.4}$$

Now it is convenient to define quantities expressing power, conductance and susceptance in dimensionless form.

$$p = X^2G_L/2.5N^2y_e \tag{9.5}$$

$$G_1 = G_L/N^2y_e \tag{9.6}$$

$$B_1 = B_L/N^2y_e. \tag{9.7}$$

The power P produced by the electron stream and dissipated in G_L is related to p

$$P = \left(\frac{2.5I_0V_0}{\theta}\right) p. \tag{9.8}$$

In terms of p and G_1, (9.3) can be written

$$p = (1/1.25)(2.5p/G_1)^{-\frac{1}{2}}J_1[(2.5p/G_1)^{\frac{1}{2}}] \cos \Delta\theta. \tag{9.9}$$

By dividing (9.4) by (9.3), we obtain

$$\Delta\omega/\omega_0 = (-G_L/2N^2M) \tan \Delta\theta - B_L/2N^2M \tag{9.10}$$

$$-(2M/y_e)\Delta\omega/\omega_0 = G_1 \tan \Delta\theta + B_1. \tag{9.11}$$

[8] To avoid confusion on the reader's part, it is perhaps well to note that we are, for the sake of generality, changing nomenclature. Hitherto we have used Y_L to denote the load at the oscillator. Actually our load as the appendix shows is usually coupled by some transformer whose equivalent transformation ratio is $1/N^2$, so that the admittance at the gap will be Y_L/N^2.

Equations (9.9) and (9.11) give the behavior of a reflex oscillator with zero output circuit loss as the load is changed. It is interesting to plot this behavior on a Smith chart.[9] Such a plot is known as a Rieke diagram or an impedance performance chart. Suppose we first make a plot for $\Delta\theta = 0$. This is shown in Fig. 32. Constant p contours are solid and, as

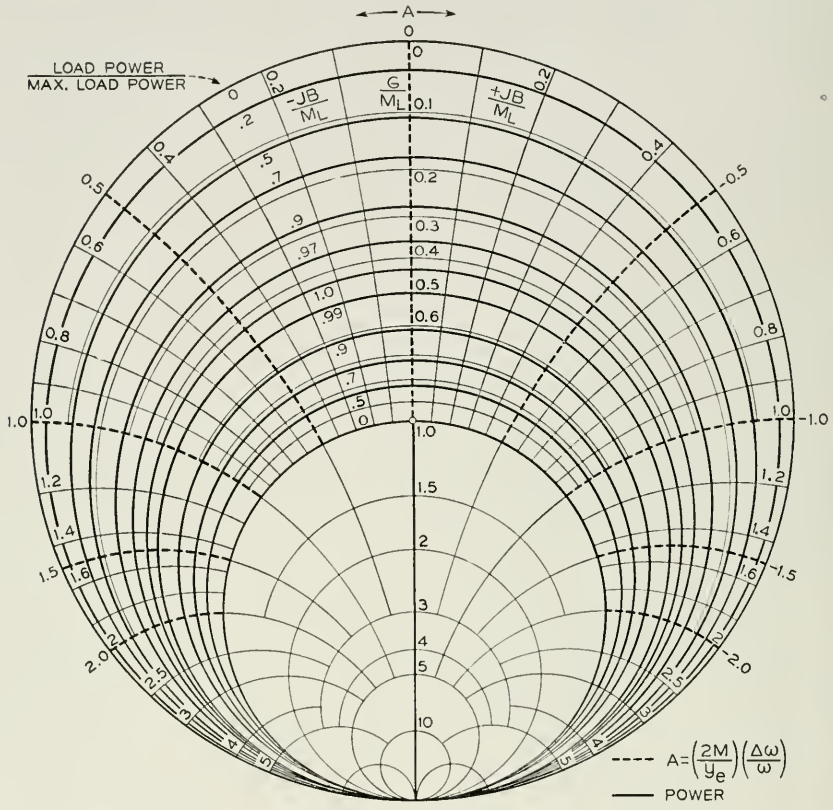


Fig. 32.—Theoretical Rieke diagram for a reflex oscillator operating with optimum drift angle. The resonator is assumed lossless. Admittances are normalized in terms of the small signal electronic admittance of the oscillator so that oscillation will stop for unity standing wave.

can be seen from the above, they will coincide with the constant conductance lines of the chart. Constant frequency curves are dashed and, for $\Delta\theta = 0$, they coincide with the locii of constant susceptance. The numbers on the frequency contours give values of $(2M/y_e)(\Delta\omega/\omega_0)$. The choice of units is such that $G_1 = 1$ means that the load conductance is just equal to small signal electronic conductance which, it will be recalled, is the starting condition for oscillation. Hence, the $G_1 = 1$ contour is a zero power contour. Any larger values of G_1 will not permit oscillation to start, so the G_1 contour

[9] P. H. Smith, "Transmission Line Calculator," *Electronics*, Jan. 1939, pp. 29-31

bounds a region of zero power commonly called the "sink," since all the frequency contours converge into it. The other zero power boundary is the outer boundary of the chart, $G_1 = 0$, which, of course, is an open circuit load. The power contours on this chart occur in pairs, except the maximum power contour which is single. These correspond to coupling greater than and less than the optimum.

The value of G_1 for any given power contour for $\Delta\theta = 0$ may be determined by referring to Fig. 9. We are assuming no resonator loss so we use the curve for which $G_R/y_e = 0$. From (9.5), if $p = 1$ we have $G_L/N^2y_e = 2.5/X^2$ which, substituted in (9.3), gives $XJ_1(X) = 1.25$. This is just the condition for maximum power output with no resonator loss. From this it can be seen that we have chosen a set of normalized coordinates. Hence, in using Fig. 9, we have $p = H/H_m$, where $H_m = .394$ is the maximum generalized efficiency. Thus, for any given value of p we let H in Fig. 9 have the value $.394p$ and determine the two values of G_1 corresponding to that contour.

From Fig. 32 we can construct several other charts describing the performance of reflex oscillators under other conditions. For instance, suppose we make $\Delta\theta$ other than zero. Such a condition commonly occurs in use either through erroneous adjustment of the repeller or through intentional use of the electronic tuning of the oscillator. We can construct a new chart for this condition using Fig. 32. Consider first the constant power contours. Suppose we consider the old contour of value p_n lying along a conductance line G_{1n}. To get a new contour, we can change the label from p_n to $p_m = p_n \cos\Delta\theta$, and we move the contour to a conductance line $G_n = G_m \cos\Delta\theta$. That this is correct can be seen by substituting these values in (9.9). Consider a given frequency contour lying along B_1. We shift each point of this contour along a constant conductance line G_{1n} an amount $B_m = G_{1n} \tan\Delta\theta$. It will be observed that this satisfied (9.11). In Fig. 33 this has been done for $\tan\Delta\theta = 1$, $\cos\Delta\theta = \sqrt{2}/2$.

Now let us consider the effect of resonator loss. Suppose we have a shunt resonator conductance G_R. Let

$$G_2 = G_R/y_e. \tag{9.12}$$

Then, if the total conductance is G_n, the fraction of the power produced which goes to the load is

$$f = (G_n - G_2)/G_n = G_1/(G_1 + G_2) \tag{9.13}$$

accordingly, we multiply each power contour label by the fraction f. Then we move all contour points along constant susceptance lines to new values

$$G_m = G_n - G_2 \tag{9.14}$$

In Fig. 34, this has been done to the contours of Fig. 32, for $G_2 = .3$.

The diagrams so far obtained have been based on the assumption that $\Delta\theta$ has been held constant. To obtain such a diagram experimentally would be extremely difficult. It would require that, as the frequency changed through load pulling, and hence the total transit angle $\theta = 2\pi f\tau$ changed, an adjustment of the repeller voltage be made to correct the change. In actual practice, Rieke diagrams for a reflex oscillator are usually made holding the

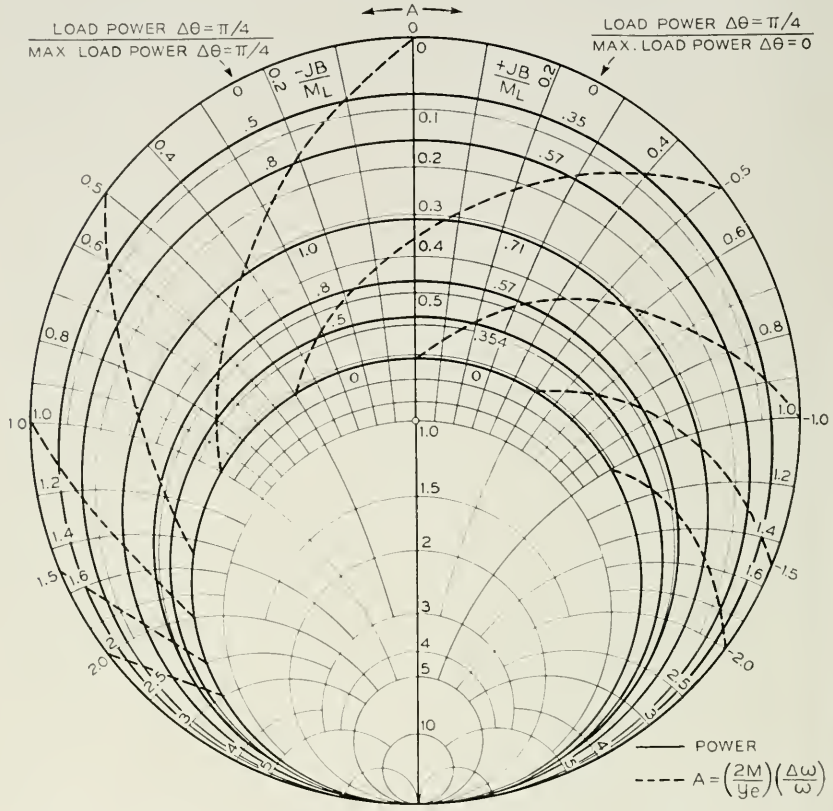


Fig. 33.—A transformation of the Rieke diagram of Fig. 32 showing the effect of shifting the drift angle away from the optimum by 45°.

transit time τ constant or in other words, with fixed operating voltages. What this does to the basic diagram of Fig. 32 is not difficult to discover, provided that $\delta\theta$ is sufficiently small so that we may ignore the variations of the Bessels functions with $\delta\theta$. We will first investigate the effect of fixed repeller voltage on the constant frequency contours. To do this we will rewrite (9.11), replacing $\Delta\theta$ by $\Delta\theta + \delta\theta$ and expand.

$$\delta\theta = \Delta\omega\tau = \frac{\Delta\omega}{\omega_0}\omega_0\tau \tag{9.15}$$

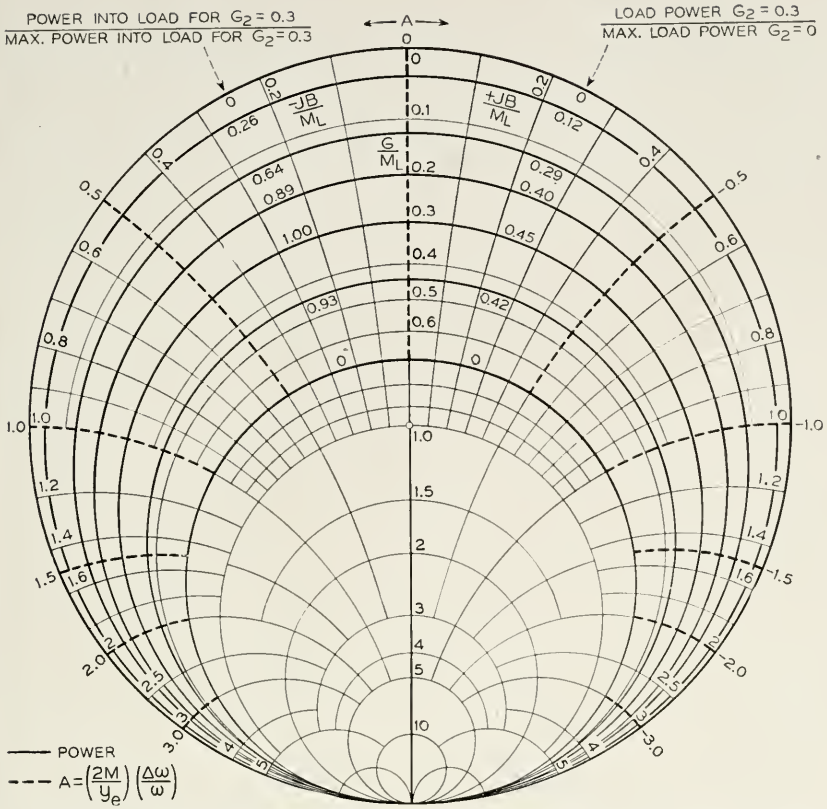


Fig. 34.—A transformation of the Rieke diagram of Fig. 32 to show the effect of the resonator loss if the phase angle is assumed to be optimum.

In rewriting (9.11) we will also replace G_1 by $G_1 + G_2$, to take resonator loss into account. We obtain for very small values of $\delta\theta$

$$-(2M/y_e)(\Delta\omega/\omega_0) = ((G_1 + G_2)\tan\Delta\theta + B_1)S \tag{9.16}$$

$$S = 1/(1 + (G_1 + G_2)\omega_0\tau/(2M/y_e)\cos^2\Delta\theta)$$

$$S = 1/(1 + \omega_0\tau/2Q\cos^2\Delta\theta). \tag{9.17}$$

Q is the loaded Q of the oscillator.

To obtain the new constant frequency contours in the case of $\Delta\theta = 0$ we shift each point of the old contour from its original position at a susceptance B_n along a constant conductance line G_n to a new susceptance line $B_m = B_n/S$. This neglects a second order correction. It will be observed that for small values of the conductance G_1 near the outer boundary, the frequency shifts will be practically unchanged, but near the sink where the

conductance G_1 is large the effect is to shift the constant frequency contours along the sink boundary away from the zero susceptance line to larger susceptance values. Hence, the constant frequency contours no longer coincide with the constant susceptance contours, not even for $\Delta\theta = 0$.

The change in the power contours is considerably more marked. As the frequency of the oscillator changes the transit angle is shifted from the optimum value by an amount $\delta\theta = (\Delta\omega/\omega_0)\omega_0\tau$. Thus the electronic conductance is reduced in magnitude by a factor $\cos \frac{\Delta\omega}{\omega_0}\omega_0\tau$. In particular, for the sink contour where the load conductance is just equal to the electronic conductance we see that when the repeller voltage is held constant the 0 power contour lies not on the $G_1 = 1 - G_2$ contour but on the locus of values $G_1 = \cos \frac{\Delta\omega}{\omega_0}\omega_0\tau - G_2$.

In order to determine the power contours when the transit time rather than the transit angle is held constant we make use of (9.3) with addition of resonator loss. In normalized coordinates ((9.6) and (9.12)) and for a phase angle of electronic admittance $\delta\theta$ we have

$$G_1 + G_2 = \frac{2J_1(X)}{X}\cos\delta\theta . \tag{9.18}$$

From (9.5) and (9.13) we have for the power output

$$p = \frac{G_1}{G_1 + G_2}\frac{2XJ_1(X)}{2.5}\cos\delta\theta . \tag{9.19}$$

Along any constant frequency contour $\delta\theta$ is constant and has the value given by (9.15) in terms of ω_0 and $\omega_0\tau$. Hence, it will be convenient to plot $(G_1 + G_2)$ vs X for various values of $\delta\theta$ as a parameter. This has been done in Fig. 35. The angle $\delta\theta$ has been specified in terms of a parameter A which appears in the Rieke diagrams as a measure of frequency deviation.

$$A = \frac{2M}{y_e}\frac{\Delta\omega}{\omega_0} \tag{9.20}$$

In terms of the parameter A

$$\delta\theta = (y_e/2M)(\omega_0\tau)A . \tag{9.21}$$

Once we have the curves of Fig. 35 we can find the power for any point on the impedance performance chart. We may, for instance, choose to find the power along the constant frequency contours, for each of which A (or $\delta\theta$) has certain constant values. We assume some constant resonator loss G_2. Choosing a point along the contour is merely taking a particular value of G_1. Having $\delta\theta$, G_2 and G_1 we can obtain X from Fig. 35. Then, knowing X, we can calculate the power from (9.19).

In constructing an impedance performance chart we want constant power contours. In obtaining these it is convenient to assume a given value of G_2. We will use $G_2 = .3$ as an example. Then we can use Fig. 35 and

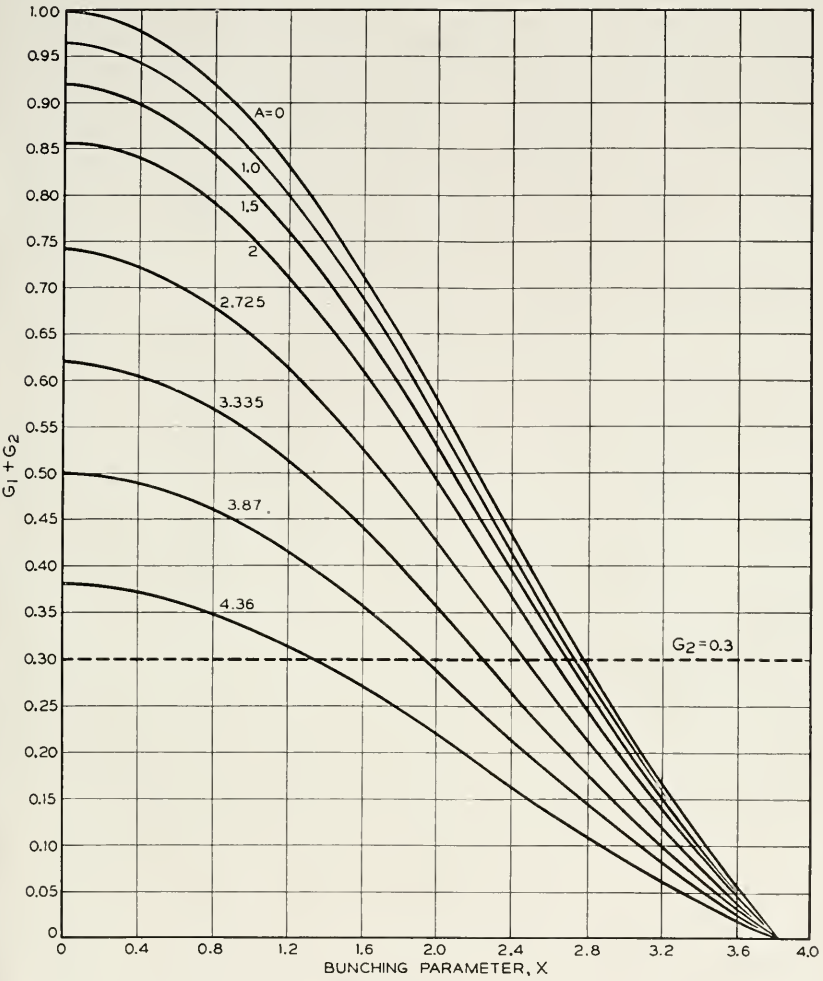


Fig. 35.—Curve of load plus loss conductance vs bunching parameter X for various values of a parameter A which gives the deviation in the drift time from the optimum time. The load and loss conductance are normalized in terms of the small signal electronic admittance. The horizontal line represents a loss conductance of $G_2 = .3$.

(9.19) to construct a family of curves giving p vs G_1 with A (or $\delta\theta$) as a parameter. In a particular case it was assumed that

$$M/y_e = 90$$

$$\omega_0\tau = 2\pi(7 + \tfrac{3}{4}).$$

These values are roughly those for the 2K25 reflex oscillator. Figure 36 shows p vs G_1 for the particular parameters assumed above. The curves were obtained by assuming values of G_1 for an appropriate A and so obtaining values of X from Fig. 35. Then the power was calculated using (9.19) and so a curve of power vs G_1 for a particular value of A was constructed.

Figure 37 shows an impedance performance chart obtained from (9.16) and Fig. 36. In using Fig. 36 to obtain constant power contours, we need merely note the values of G_1 at which a horizontal line on Fig. 36 intersects the curves for various values of A. Each curve either intersects such a horizontal (constant power) line at two points, or it is tangent or it does not intersect. The point of tangency represents the largest value of A at which the power can be obtained, and corresponds to the points of the crescent shaped power contours of the impedance performance chart. The maximum power contour contracts to a point.

Along the boundary of the sink, for which $p = 0$, $X = 0$ and we have from (9.18)

$$G_1 = \cos \delta\theta - G_2 . \tag{9.22}$$

The results which we have obtained can be extended to include the case in which $\Delta\theta \neq 0$. Further, as we know from Appendix I, we can take into account losses in the output circuit by assuming a resistance in series with the load. In a well-designed reflex oscillator the output circuit has little loss. The chief effect of this small loss is to round off the points of the constant power contours.

In actually measuring the performance of an oscillator, output and frequency are plotted vs load impedance as referred to the characteristic impedance of the output line. Also, frequently the coupling is adjusted so that for a match (the center of the Smith chart) optimum power is obtained. We can transform our impedance performance chart to correspond to such a plot by shifting each point G, B on a contour to a new point

$$G_1 = G/G_{\max}$$

$$B_1 = B/G_{\max}$$

where $G_{\max}$ is the conductance for which maximum power is obtained. Such a transformation of Fig. 37 is shown in Fig. 38.

It will be noted in Fig. 38 that the standing wave ratio for 0 power, the sink margin, is about 2.3. This sink margin is nearly independent of the resonator loss for oscillators loaded to give maximum power at unity standing wave ratio, as has been discussed and illustrated in Fig. 10. If the sink margin must be increased or the pulling figure must be decreased[10] the coup-

[10] The pulling figure is arbitrarily defined as the maximum frequency excursion produced when a voltage standing wave ratio of $\sqrt{2}$ is presented to the oscillator and the phase is varied through 180°.

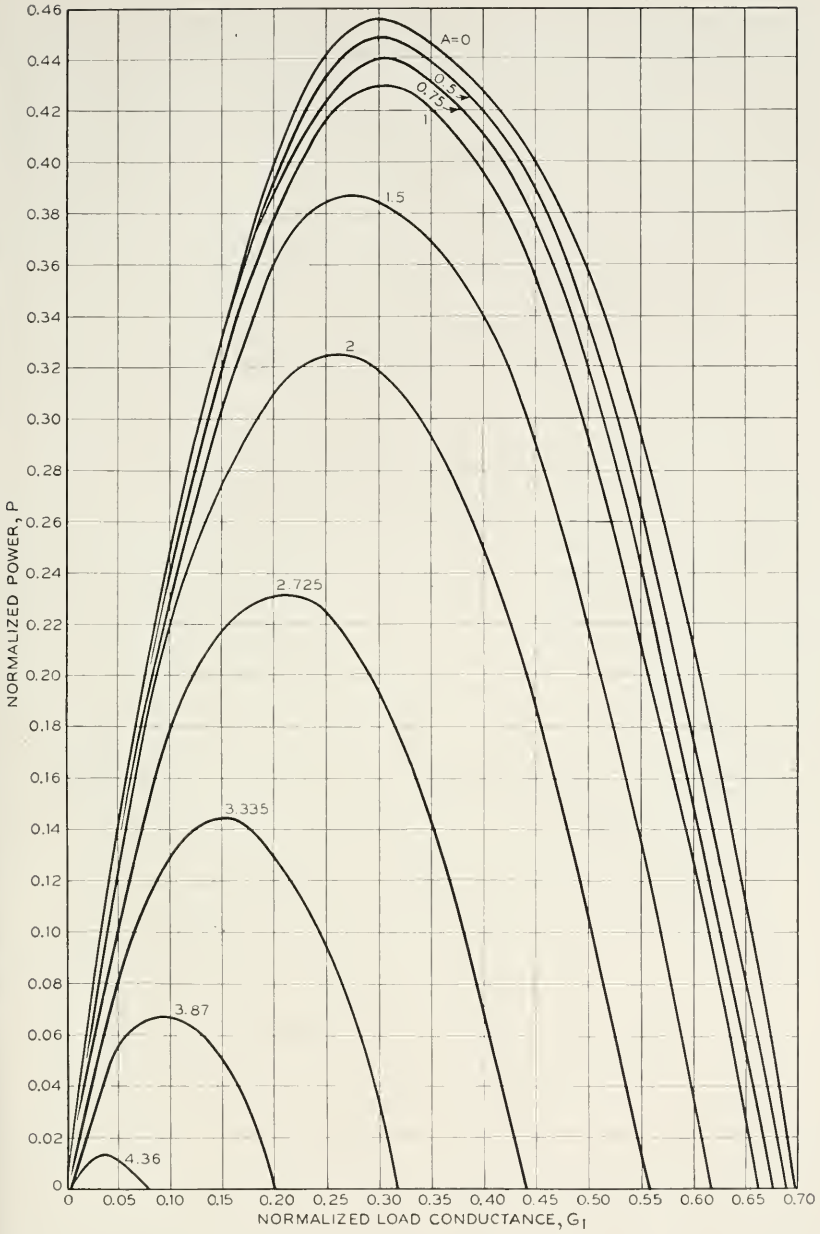


Fig. 36.—Normalized power vs normalized load conductance for various values of the parameter A which gives the deviation in drift time from the optimum drift time. These curves are computed for the case $G_2 = .3$. Optimum drift angle equal to 15.5 π radians and a ratio of characteristics resonator admittance to small signal electronic admittance of 90 is assumed.

ling can be reduced so that for unity standing wave ratio the load conductance appearing at the gap is less than that for optimum power.

Finally, in making measurements the load impedance is usually evaluated at a point several wavelengths away from the resonator. If performance is plotted in terms of impedances so specified, the points on the contours of

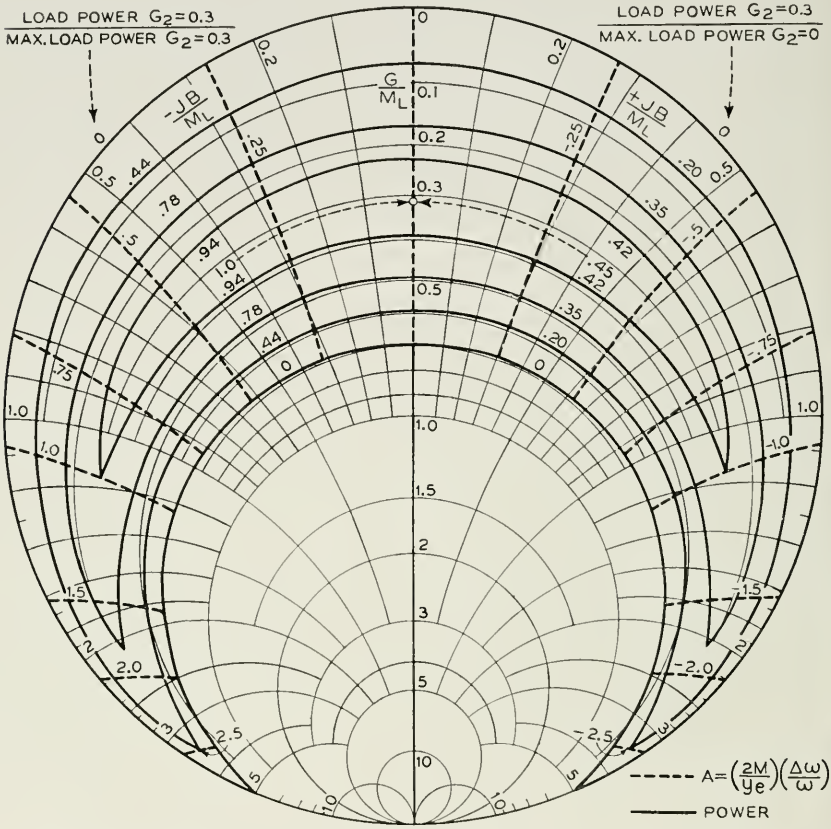


Fig. 37.—A Rieke diagram for a reflex oscillator having a lossy resonator, taking into account the variation of drift angle with frequency pulling. This results in closed power contours.

Fig. 38 appear rotated about the center. As the line length in wavelengths will be different for different frequencies, points on different frequency contours will be rotated by different amounts. This can cause the contours to overlap in the region corresponding to the zero admittance region of Fig. 38. With very long lines, the contours may overlap over a considerable region. The multiple modes of oscillation which then occur are discussed in somewhat different terms in the following section.

Figure 39 shows the performance chart of Fig. 38 as it would appear with the impedances evaluated at a point 5 wavelengths away from the resonator. Figure 71 of Section XIII shows an impedance performance chart for 2K25 reflex oscillator.

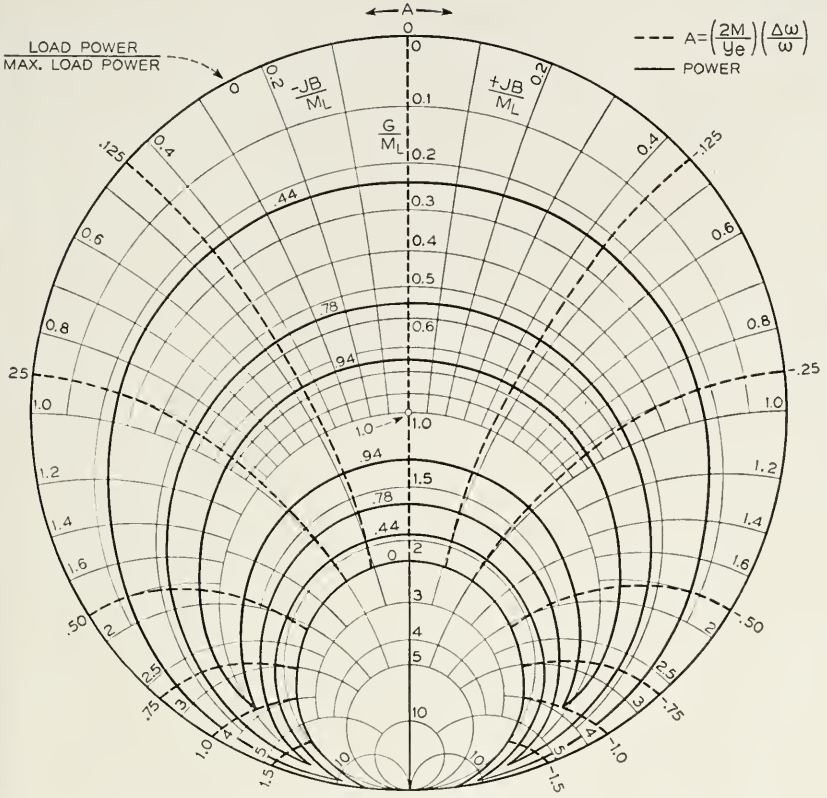


Fig. 38.—The Rieke diagram of Fig. 37 transformed to apply to the oscillator loaded for optimum power at unity standing wave.

B. *Frequency—Sensitive Loads—Long Line Effect*

When the load presented to a reflex oscillator consists of a long line mismatched at the far end, or contains a resonant element, the operation of a reflex oscillator, and especially its electronic tuning, may be very seriously affected.

For instance, consider the simple circuit shown in Fig. 40. Here M_R is the characteristic admittance of the reflex oscillator resonator as seen from the output line or wave guide and M_L is the characteristic impedance of a line load ℓ long, so terminated as to give a standing wave ratio, σ.

In the simple circuit assumed there are essentially three variables; (1) the ratio of the characteristic admittance of the resonant circuit, M_R to

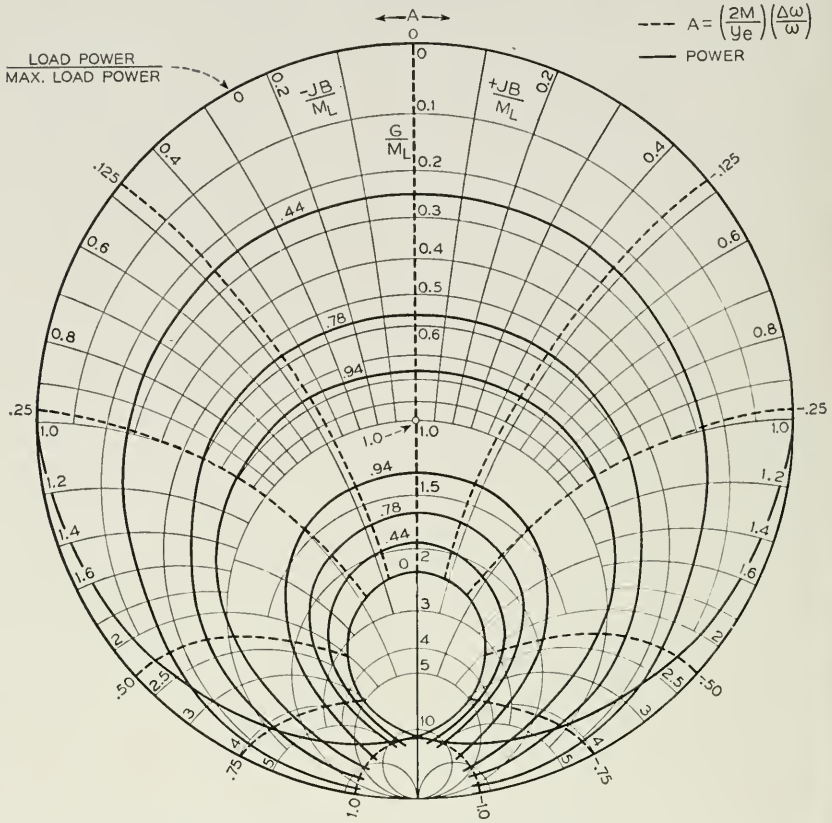


Fig. 39.—The Rieke diagram of Fig. 38 transformed to include the effect of a line five wave lengths long between the load and the oscillator.

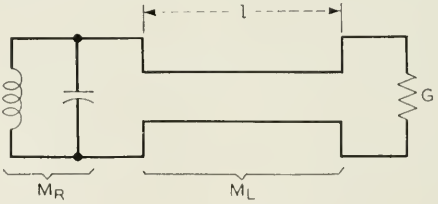


Fig. 40.—Equivalent circuit of a lossless resonator, a line and a mismatched load.

that of the line, M_L. This ratio will be called the external Q and signified by Q_E

$$Q_E = M_R/M_L . \tag{9.23}$$

For a lossless resonator and unity standing wave ratio, the loaded Q is equal to Q_E. For a resonator of unloaded Q, Q_0, and for unity standing wave ratio, the loaded Q, obeys the relation

$$1/Q = 1/Q_E + 1/Q_0 \tag{9.24}$$

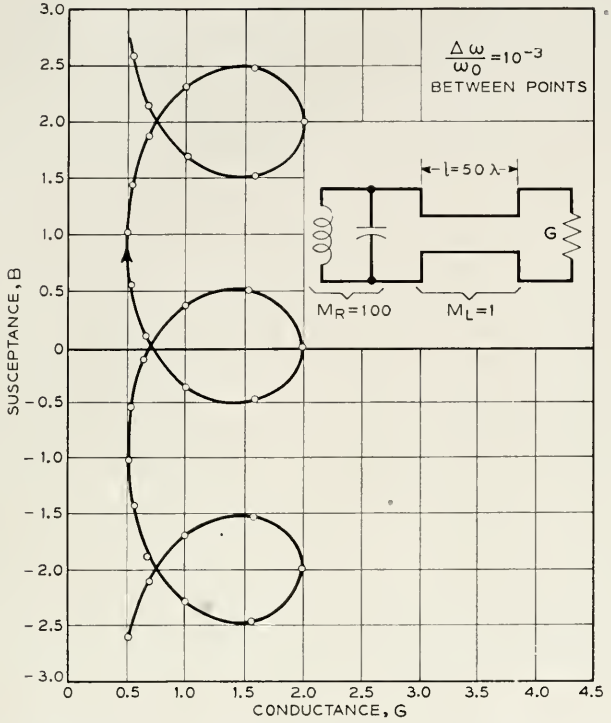


Fig. 41.—Susceptance vs conductance for a resonator coupled to a 50 wave length line terminated by a load having a standing wave ratio of 2. Characteristic admittance of the resonator is assumed to be equal to 100 in terms of a line characteristic admittance of unity. The circles mark off relative frequency increments

$$\frac{\Delta\omega}{\omega_0} = 10^{-3},$$

where ω_0 is the frequency of resonance.

(2) the length of the line called θ when measured in radians or n when measured in wavelengths, (3) the standing wave ratio σ.

Figures 41 and 42 show admittance plots for two resonant circuits loaded by mismatched lines of different lengths. The feature to be observed is the loops, which are such that at certain points the same admittance is achieved at two different frequencies. It is obvious that a line representing $-Y_e$

may cut such a curve at more than one point: thus, oscillation at more than one frequency is possible. Actually, there may be three intersections per loop. The two of these for which the susceptance B is increasing with frequency represent stable oscillation; the intersection at which B is decreasing with frequency represents an unstable condition.

The loops are of course due to reactance changes associated with variation of the electrical length of the line with frequency. Slight changes in tuning of the circuit or slight changes in the length of the line shift the loops up or down, parallel to the susceptance axis. Thus, whether the electronic admittance line actually cuts a loop, giving two possible oscillating frequencies, may depend on the exact length of the line as well as on the ex-

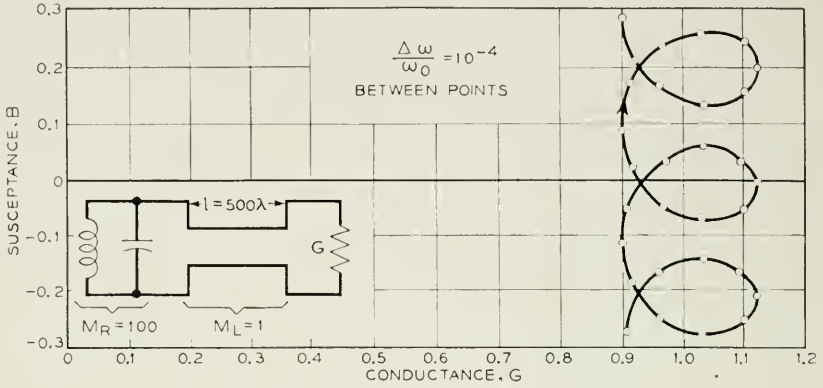


Fig. 42.—Susceptance vs conductance for line 500 wave lengths long terminated by a load having a standing wave ratio of 1.11. Circles mark off relative frequency increments of 10^{-4}. Characteristic admittance to the resonator equals 100.

istence of loops. The frequency difference between loops is such as to change the electrical length of the line by one-half wavelength.

The existence or absence of loops and their size depend on all three parameters. Things which promote loops are:

Low ratio of M_R/M_L or Q_E

Large n or θ

High σ

As any parameter is changed so as to promote the existence of loops, the Y curve first has merely a slight periodic variation from the straight line for a resistively loaded circuit. Further change leads to a critical condition in which the curve has cusps at which the rate of change of admittance with frequency is zero. If the electronic admittance line passes through a cusp,

the frequency of oscillation changes infinitely rapidly with load. Still further change results in the formation of loops. Further change results in expansion of loops so that they overlap, giving more than three intersections with the electronic admittance line.

Loops may exist for very low standing wave ratios if the line is sufficiently long. Admittance plots for low standing wave ratio are very nearly cycloidal in shape; those for higher standing wave ratios are similar to cycloids in appearance but actually depart considerably from cycloids in exact form.

By combining the expression for the near resonance admittance of a tuned circuit with the transmission line equation for admittances, the expression for these admittance curves is obtained. Assuming the termination to be an admittance Y_T which at frequency ω_0 is θ_0 radians from the resonator,

$$Y = j2M_R\Delta\omega/\omega_0 + M_L\frac{(Y_T/M_L) + j\tan\theta_0(1 + \Delta\omega/\omega_0)}{1 + j(Y_T/M_L)\tan\theta_0(1 + \Delta\omega/\omega_0)}. \quad (9.25)$$

The critical relation of parameters for which a cusp is formed is important, for it divides conditions for which oscillation is possible at one frequency only and those for which oscillation is possible at two frequencies. This cusp corresponds to a condition in which the rate of change with frequency of admittance of the mismatched line is equal and opposite to that of the circuit. This may be obtained by letting Y_T be real.

$$Y_T/M_L > 1, \qquad \theta_0 = n\pi \text{ where } n \text{ is an integer.}$$

The standing wave ratio is then

$$\sigma = Y_T/M_L. \quad (9.26)$$

The second term on the right of (9.25) is then

$$Y_2 = M_L\left(\frac{\sigma + j\tan\theta_0\Delta\omega/\omega_0)}{1 + j\sigma\tan\theta_0\Delta\omega/\omega_0}\right). \quad (9.27)$$

For very small values of $\Delta\omega$ we see that very nearly

$$Y_2 = M_L[\sigma - j(\sigma^2 - 1)\theta_0\Delta\omega/\omega_0]. \quad (9.28)$$

Thus for the rate of change of total admittance to be zero

$$\begin{aligned} 2M_R &= M_L(\sigma^2 - 1)\theta_0 \\ \theta_0 &= 2(M_R/M_L)(\sigma^2 - 1) \\ &= 2Q_E/(\sigma^2 - 1). \end{aligned} \quad (9.30)$$

Thus, the condition for no loops, and hence, for a single oscillating frequency, may be expressed

$$\theta_0 < 2Q_E/(\sigma^2 - 1) \quad (9.31)$$

or

$$\sigma < \sqrt{1 + 2Q_E/\theta_0}.$$

We will remember that θ_0 is the length of line in radians, σ is the standing wave ratio, measured as greater than unity, and Q_E is the external Q of the resonator for unity standing wave ratio.

Replacing a given length of line by the same length of wave guide, we find that the phase angle of the reflection changes more rapidly with frequency, and instead of (9.31) we have the condition for no loops as

$$\theta < 2Q_E(1 - (\lambda/\lambda_0)^2)/(\sigma^2 - 1) \tag{9.32}$$

$$\sigma < \sqrt{1 + 2Q_E(1 - (\lambda/\lambda_0)^2)/\theta_0}.$$

Here λ is the free space wavelength and λ_0 is the cutoff wavelength of the guide.

Equations (9.32) are for a particular phase of standing wave, that is, for relations of Y_T and θ_0 which produce a loop symmetrical above the C axis. Loops above the G axis are slightly more looped than loops below the G axis because of the increase of θ_0 with frequency. For reasonably long lines, (9.32) applies quite accurately for formation of loops in any position; for short lines loops are of no consequence unless they are near the G axis.

An imporant case is that in which the resonant load is coupled to the resonator by means of a line so short that it may be considered to have a constant electrical length for all frequencies of interest. The resonant load will be assumed to be shunted with a conductance equal to the characteristic admittance of the line. As the multiple resonance of a long mismatched line resulted in formation of many loops, so in this case we would rightly suspect the possibility of a single loop.

If the resonant load is $\frac{1}{4}$, $\frac{3}{4}$, etc. wavelengths from the resonator, and both resonate at the same frequency, a loop is formed symmetrical about the G axis. Figure 43 is an admittance curve for resonator and load placed $\frac{1}{4}$ wavelength apart. Tuning either resonator or load moves this loop up or down.

If the distance from resonator to resonant load is varied above or below a quarter wave distance, the loop moves up or down and expands. This is illustrated by an eighth wavelength diagram for the same resonator and load as of Fig. 43 shown in Fig. 44.

When the distance from the resonator to the resonant load, including the effective length of the coupling loop, is $\frac{1}{2}$, 1, $1\frac{1}{2}$, etc. wavelengths, for frequencies near resonance the resonant load is essentially in shunt with the resonator, and its effect is to increase the loaded Q of the resonator. An admittance curve for the case is shown in Fig. 45. In this case the loops

have moved considerably away in frequency, and expanded tremendously. There are still recrossings of the axis near the origin, however, as indicated in this case by the dashed line which represents 2 crossings, in this case about 4% in frequency above and below the middle crossing if the length of the line ℓ is $\lambda/2$.

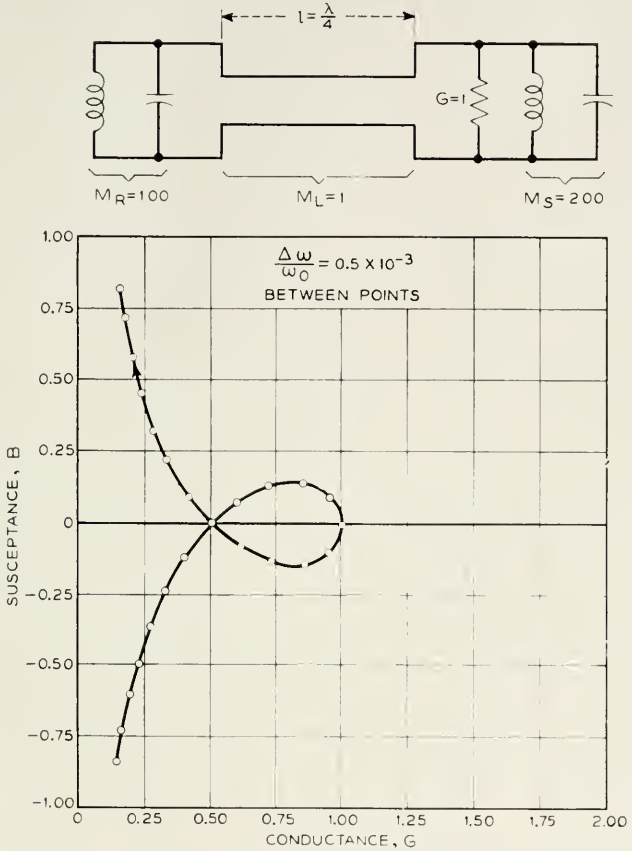


Fig. 43.—Susceptance vs conductance for two resonators coupled by a quarter wave line. The resonator at which the admittance is measured has a characteristic admittance of 100 in terms of a line characteristic admittance of unity. The other resonator has a characteristic admittance of 200 and a shunt conductance of unity. The circles mark off relative frequency increments of $\frac{1}{2} \times 10^{-3}$ in terms of the resonant frequency.

As a sort of horrible example, an admittance curve for a high Q load 50 wavelengths from the resonator was computed and is shown in Fig. 46. Only a few of the loops are shown.

Admittance curves for more complicated circuits such as impedance transformers can be computed or obtained experimentally.

As has been stated, one of the most serious effects of such mismatched long line or resonant loads is that on the electronic tuning. For instance, consider the circuit admittance curve to be that shown in Fig. 47, and the minus electronic admittance curve to be a straight line extending from the origin. As the repeller voltage is varied and this is swung down from the $+B$ axis its extreme will at some point touch the circuit admittance line

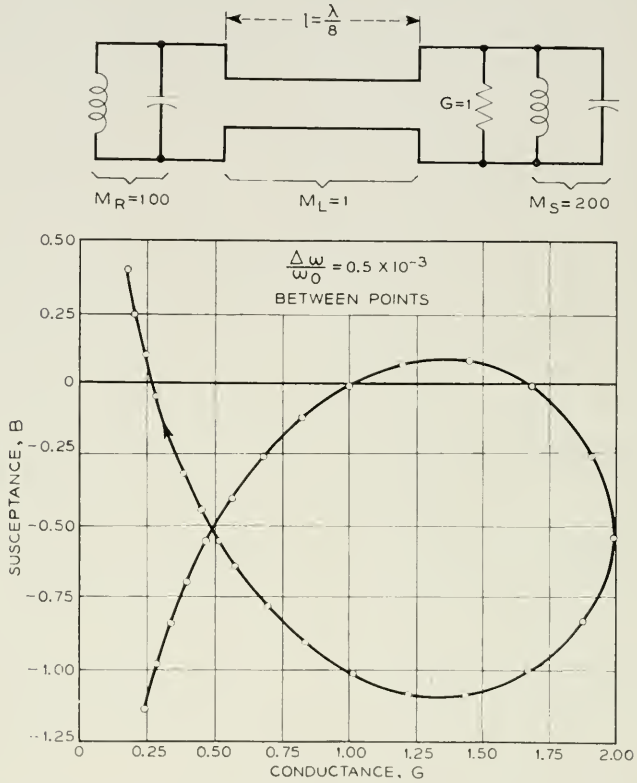


Fig. 44.—Susceptance vs conductance for the same resonators as of Fig. 43 coupled by a one-eighth wave line.

and oscillation will commence. As the line is swung further down, the frequency will decrease. Oscillation will increase in amplitude until the $-Y_e$ line is perpendicular to the Y line. From that point on oscillation will decrease in amplitude until the $-Y_e$ line is parallel to the Y curve on the down side of the loop. Beyond this point the intersection cannot move out on the loop, and the frequency and amplitude will jump abruptly to correspond with the other intersection. As the $-Y_e$ line rotates further,

amplitude will decrease and finally go to zero when the end of the $-Y_e$ line touches the Y curve. If the $-Y_e$ line is rotated back, a similar phenomenon is observed. This behavior and the resulting electronic tuning characteristic are illustrated in Figs. 47 and 48. Such electronic tuning

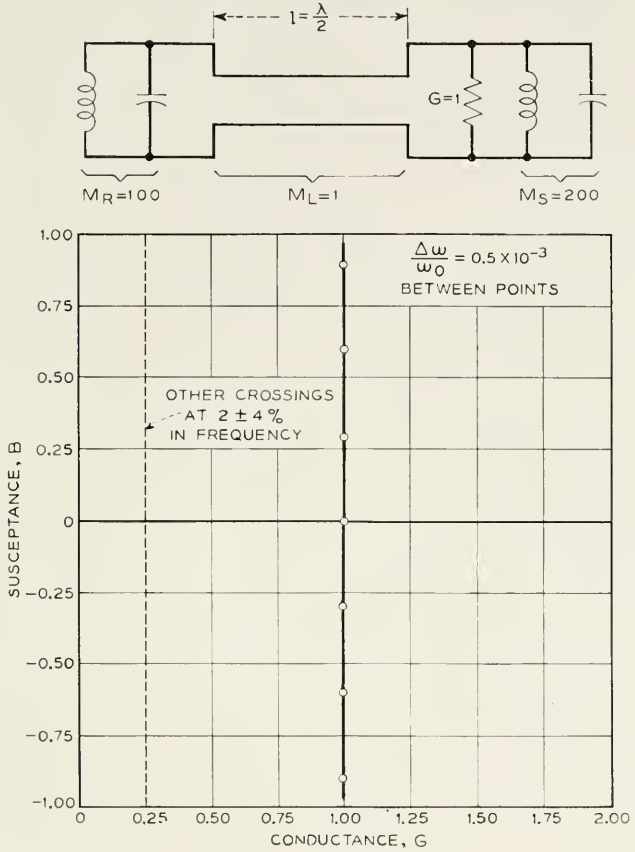


Fig. 45.—Susceptance vs conductance for the same resonators as of Fig. 43 coupled by a one-half wave line. The dash line indicates two other crossings of the 0 susceptance axis, at frequencies ±4% from the resonant frequency of the resonators.

characteristics are frequently observed when a reflex oscillator is coupled tightly to a resonant load.

C. *Effect of Short Mismatched Lines on Electronic Tuning*

In the foregoing, the effect of long mismatched lines in producing additional multiple resonant frequencies and possible modiness in operation has

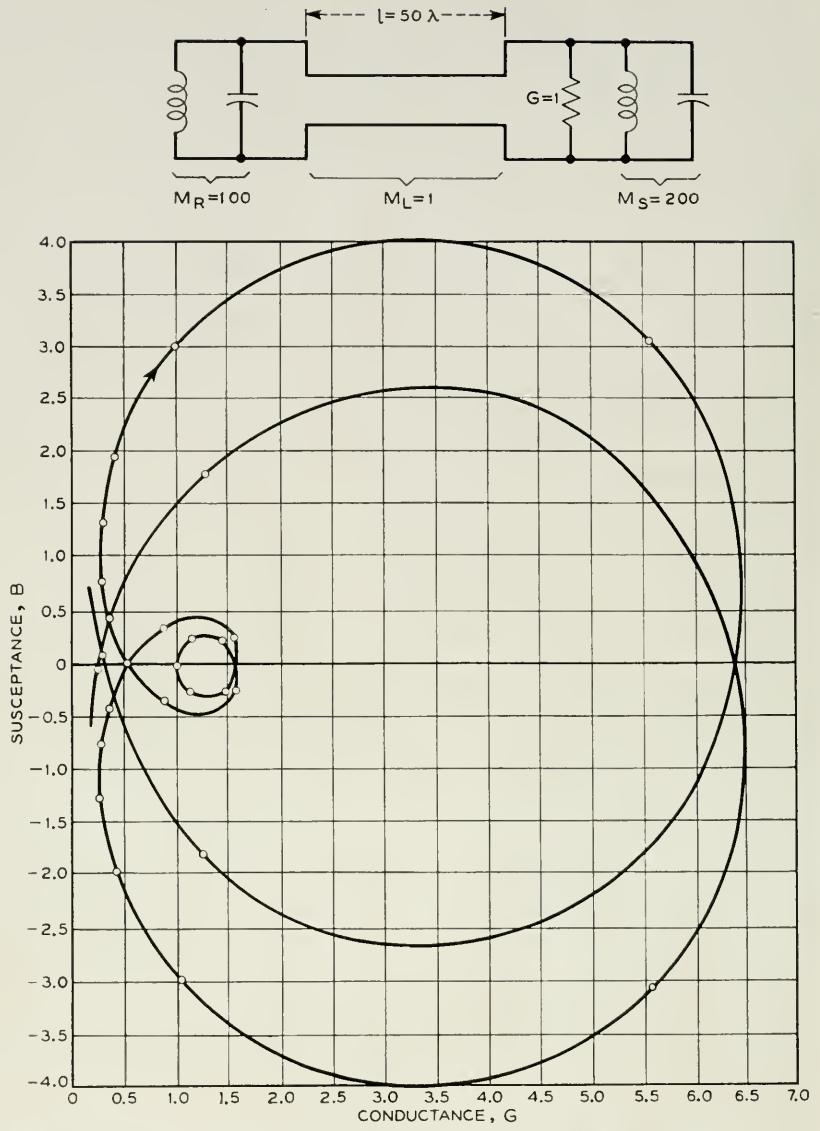


Fig. 46.—Susceptance vs conductance for the resonators of Fig. 43 coupled by a line 50 wave lengths long.

been explained. The effect of such multiple resonance on electronic tuning has been illustrated in Fig. 48.

If a short mismatched line is used as the load for a reflex oscillator, there

may be no additional modes, or such modes may be so far removed in frequency from the fundamental frequency of the resonator as to be of little

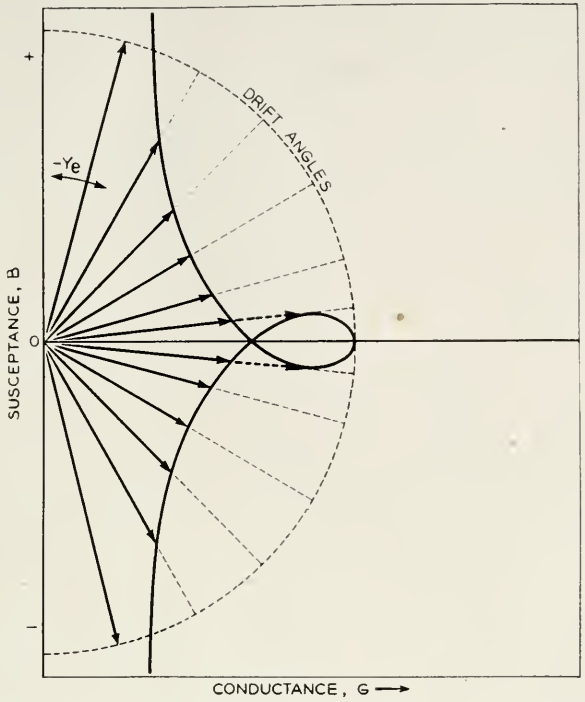


Fig. 47.—Behavior of the intersection between a circuit admittance line with a loop and the negative of the electronic admittance line of a reflex oscillator as the drift angle is varied (circuit hysteresis).

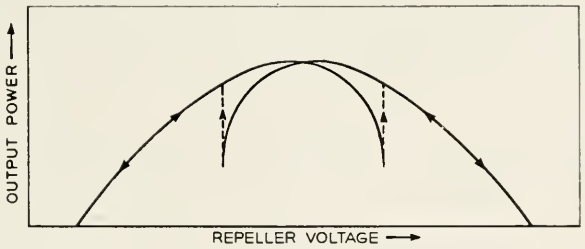


Fig. 48.—Output vs repeller voltage for the conditions obtaining in Fig. 47.

importance. Nonetheless, the short line will add a frequency-sensitive reactance in shunt with the resonant circuit, and hence will change the characteristic admittance of the resonator.

Imagine, for instance, that we represent the resonator and the mismatched line as in shunt with a section of line N wavelengths or θ radians long misterminated in a frequency insensitive manner so as to give a standing wave ratio σ. If M_L is the characteristic admittance of the line, the admittance it produces at the resonator is

$$Y_L = M_L \frac{\sigma + j \tan \theta}{1 + j\sigma \tan \theta}. \tag{9.33}$$

Now, if the frequency is increased, θ is made greater and Y is changed.

$$\partial Y_L/\partial\theta = M_L \frac{j(1 - \sigma^2) \sec^2 \theta}{(1 + j\sigma \tan \theta)^2}. \tag{9.34}$$

We are interested in the susceptive component of change. If

$$Y_L = G_L + jB_L \tag{9.35}$$

we find

$$\partial B_L/\partial\theta = M_L \frac{(1 - \sigma^2)(1 - \sigma^2 \tan^2 \theta) \sec^2 \theta}{(1 + \sigma^2 \tan^2 \theta)}. \tag{9.36}$$

Now, if frequency is changed by an amount df, θ will increase by an a mount $\theta(df/f)$ and B_L will change by an amount

$$dB_L = (\partial B_L/\partial\theta)(2\pi N)(df/f). \tag{9.37}$$

We now define a parameter M_M expressing the effect of the mismatch as follows

$$\pi(\partial B_L/\partial\theta) = M_M. \tag{9.38}$$

Then

$$dB_L = 2NM_M(df/f). \tag{9.39}$$

If the characteristic admittance of the resonator is M_R, then the characteristic admittance of the resonator plus the line is

$$M = M_R + NM_M. \tag{9.40}$$

If, instead of a coaxial line, a wave guide is used, and λ_0 and λ are the cutoff and operating wavelengths, we have

$$dB_L = 2NM_M(df/f)(1 - (\lambda/\lambda_0)^2)^{-\frac{1}{2}} \tag{9.41}$$

and

$$M = M_R + NM_M(1 - (\lambda/\lambda_0)^2)^{-\frac{1}{2}} \tag{9.42}$$

In Fig. 49 contour lines for M_M constant are plotted on a Smith Chart (reflection coefficient plane). Over most of the plane M_M has a moderate

positive value tending to increase characteristic admittance and hence decrease electronic tuning. Over a very restricted range in the high admittance region M_M has large negative values and over a restricted range outside of this region M_M has large positive values.

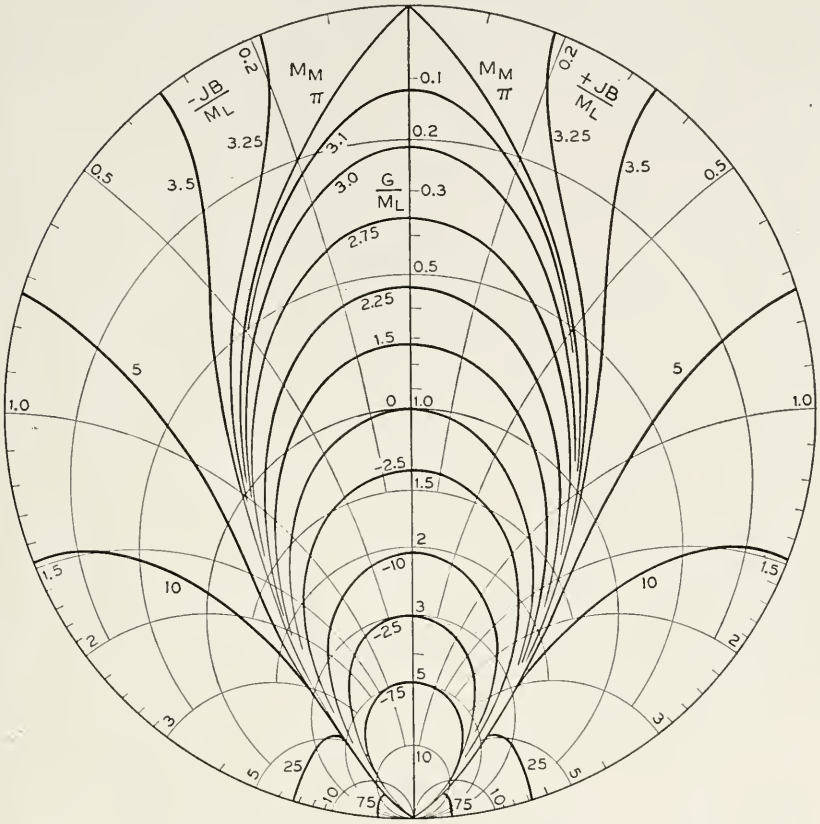


Fig. 49.—Lines of constant value of a parameter M_M shown on a chart giving the conductance and susceptance of the terminating admittance of a short line. The parameter plotted multiplied by the number of wave lengths in the line gives the additional characteristic admittance due to the resonant effects of the line. The parameter M_M is of course 0 for terminated lines (center of chart).

This is an appropriate point at which to settle the issue: what do we mean by a "short line" as opposed to a "long line." For our present purposes, a short line is one short enough so that M_M does not change substantially over the frequency range involved. Thus whether a line is short or not depends on the phase of the standing wave at the resonator (the position

on the Smith Chart) as well as on the length of the line. M_M changes most rapidly with frequency in the very high admittance region.

As a simple example of the effect of a short mismatched line on electronic tuning between half power points, consider the case of a reflex oscillator with a lossless resonator so coupled to the line that the external Q is 100 and the electronic conductance is 3 in terms of the line admittance. Suppose we couple to this a coaxial line 5 wavelengths long with a standing wave ratio $\sigma = 2$, vary the phase, and compute the electronic tuning for various

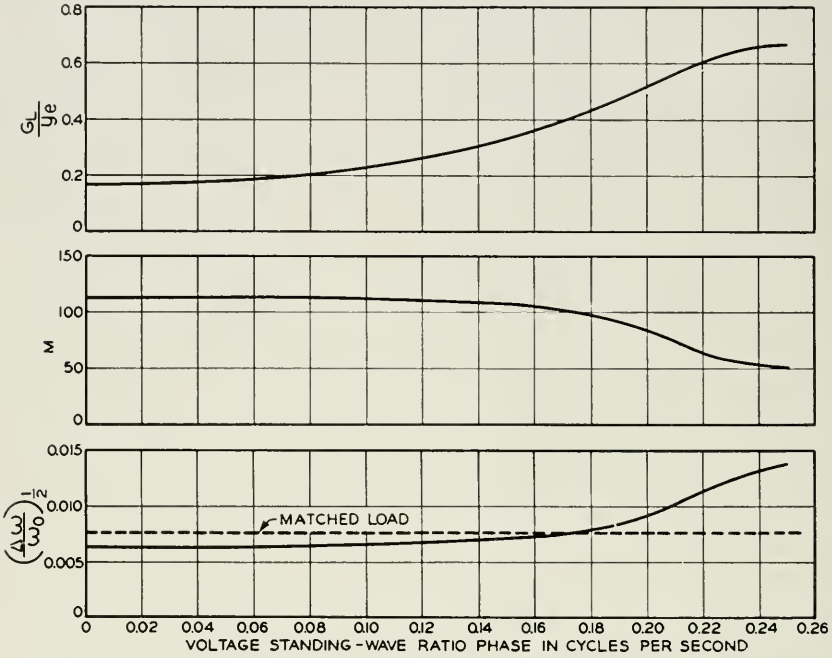


Fig. 50.—The normalized load conductance, the characteristic admittance of the resonator and the normalized electronic tuning range to half power plotted vs standing wave ratio phase for a particular case involving a short misterminated line. The electronic tuning for a matched line is shown as a heavy horizontal line in the plot of $(\Delta\omega/\omega_0)^{\frac{1}{2}}$.

phases. We can do this by obtaining the conductance and M_L from Fig. 49 and using Fig. 15 to btain $(\Delta\omega/\omega_0)^{\frac{1}{2}}$. In Fig. 50, the parameters G_L/y_e (the total characteristic admittance including the effect of the line), N, and, finally, $(\Delta\omega/\omega_0)^{\frac{1}{2}}$ have been plotted vs standing wave phase in cycles. $(\Delta\omega/\omega_0)^{\frac{1}{2}}$ for a matched load is also shown. This example is of course not typical for all reflex oscillators: in some cases the electronic tuning might be reduced or oscillation might stop entirely for the standing wave phases which produce high conductance.

X. Variation of Power and Electronic Tuning with Frequency

When a reflex oscillator is tuned through its tuning range, the load and repeller voltage being adjusted for optimum efficiency for a given drift angle, it is found that the power and efficiency and the electronic tuning vary, having optima at certain frequencies.

When we come to work out the variation of power and electronic tuning with frequency we at once notice two distinct cases: that of a fixed gap spacing and variable resonator (707A), and that of an essentially fixed resonator and a variable gap spacing (723A etc.); see Section XIII. Here we will treat as an example the latter case only.

The simplest approximation of the tuning mechanism which can be expected to accord reasonably with facts is that in which the resonator is represented as a fixed inductance, a constant shunt "stray" capacitance and a variable capacitance proportional to $1/d$, where d is the gap spacing. The validity of such a representation over the normal operating range has been verified experimentally for a variety of oscillator resonators. Let C_0 be the fixed capacitance and C_1 be the variable capacitance at some reference spacing d_1. Then, letting the inductance be L, we have for the frequency

$$\omega = (L(C_0 + C_1 d_1/d))^{-\frac{1}{2}}. \tag{10.1}$$

Suppose we choose d_1 such that

$$C_0 = C_1. \tag{10.2}$$

Then, letting

$$d/d_1 = D \tag{10.3}$$

$$\omega_1 = (2LC_0)^{-\frac{1}{2}} = 2\pi f_1 \tag{10.4}$$

$$\omega/\omega_1 = W. \tag{10.5}$$

We find

$$W = 2^{\frac{1}{2}}(1 + 1/D)^{-\frac{1}{2}}. \tag{10.6}$$

This relation is shown in Fig. 51, where D is plotted vs W. It is perfectly general (within the validity of the assumptions) for a proper choice of reference spacing d_1. We have, then, in Fig. 51 a curve of spacing D vs reduced frequency W.

The parameter which governs the power and efficency is G_R/y_e. We have

$$G_R/y_e = (G_R/\beta^2)(2V_0/I_0\theta). \tag{10.7}$$

As V_0, I_0 and θ will not vary in tuning the oscillator, we must look for variaton in G_R and β^2.

For parallel plane grids, we have

$$1/\beta^2 = (\theta_g/2)^2/\sin^2(\theta_g/2) \tag{10.8}$$

where θ_g is the transit angle between grids. We see that in terms of W and D we can write

$$\theta_g = \theta_1 WD\,. \tag{10.9}$$

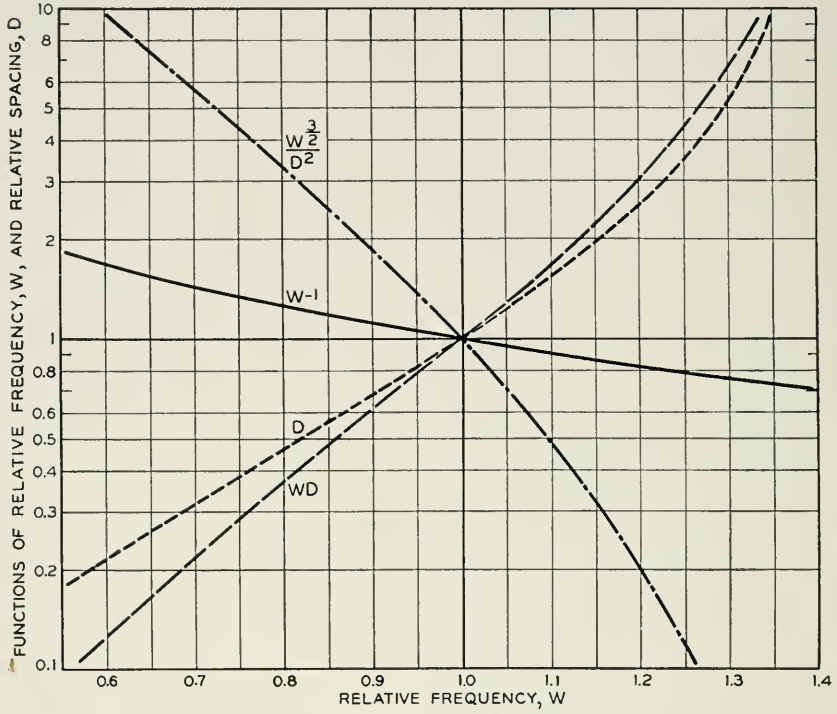


Fig. 51.—Various functions of relative frequency W and relative spacing D plotted vs relative frequency.

Here θ_1 is the gap transit angle at a spacing d_1 and a frequency W_1. So that we may see the effect of tuning on $1/\beta^2$, WD has been plotted vs W in Fig. 51 and $1/\beta^2$ has been plotted vs θ_g in Fig. 52.

We now have to consider losses. From (9.7) of Appendix IX we see that the grid loss conductance can be expressed in the form

$$G_g = G_{g1} W^{\frac{3}{2}}/D^2. \tag{10.10}$$

Here G_{g1} is the grid loss conductance at $d = d_1$ and $\omega = \omega_1$.

Finally, let us consider the resonator loss. If the resonator could be represented by an inductance L with a series resistance R, at high frequencies the conductance would be very nearly

$$G_L = R/(\omega L)^2. \tag{10.11}$$

If R varies as $\omega^{\frac{1}{2}}$, we see that we could then write

$$G_L = G_{L1} W^{-\frac{3}{2}}. \tag{10.12}$$

Here G_{L1} is the conductance at a frequency ω_1.

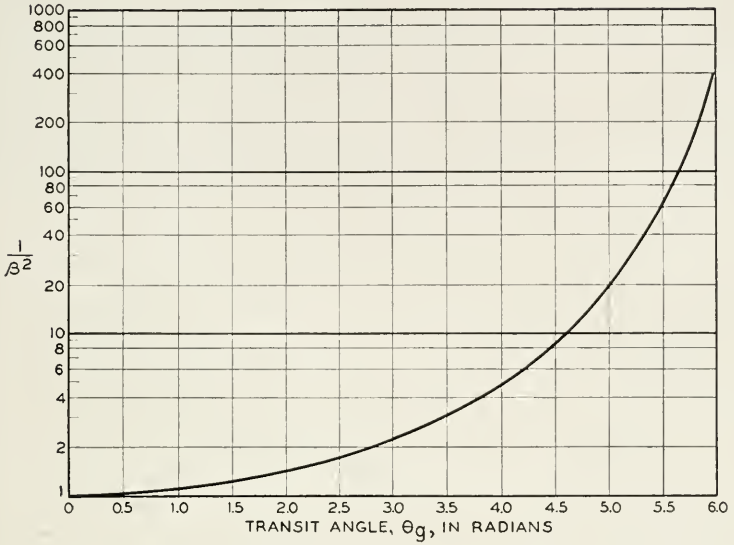


Fig. 52.—The reciprocal of the square of the modulation coefficient is a function of the gap transit angle in radians for the case of fine parallel grids.

As an opposite extreme let us consider the behaviour of the input conductance of a coaxial line. It can be shown that, allowing the resistance of such a line to vary as $\omega^{\frac{1}{2}}$, the input conductance is

$$G_\ell = A\omega^{\frac{1}{2}} \csc^2 (\omega \ell / C). \tag{10.13}$$

Here ℓ is the length of the line and C is the velocity of propagation. If G_L given by (10.12) and G_ℓ of 10.13 give the same value of conductance at some angular frequency ω_1 then it will be found that for values of ℓ typical of reflex oscillator resonators the variation of G_ℓ with ω will be significantly less than that of G_L. Although typical cavities are not uniform lines (10.13) indicates that a slower variation than (10.12) can be expected. It will be found moreover that the shape of the power output vs frequency curves are not very sensitive to the variation assumed. Hence as a reasonable compromise it will be assumed that the resonator wall loss varies as

$$G_S = G_{S1} W'^{-1}. \tag{10.14}$$

In Fig. 51 W'^{-1} has been plotted vs W'.

Now let us take an actual example. Suppose that at $D = 1$, i.e. $(d = d_1, \omega = \omega_1)$

$$\theta = 2$$

$$G_{\theta 1} = .270/y_e$$

$$G_{S1} = .095/y_e$$

The information above has been used in connection with Figs. 51 and 52 and ratio of resonator loss to small signal electronic admittance, G_R/y_e, has been plotted vs W' in Fig. 53. A 2K25 oscillator operated at a beam

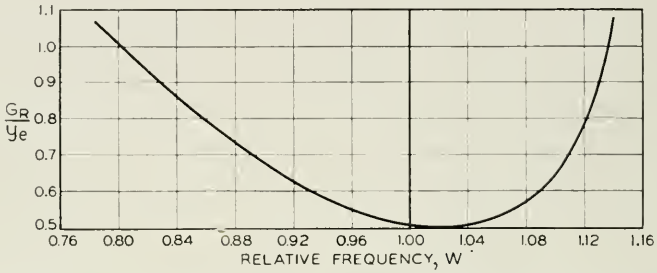


Fig. 53.—Computed variation of ratio of resonator loss to small signal electronic admittance vs relative frequency W' for certain resonator parameters assumed to fit the characteristics of the 2K25.

voltage, V_0, of 300 volts had a total cathode current I_0 of 26 ma. This current passed three grids on the first transit and back through the third grid on the return transit. On a geometrical basis, 53% of the cathode current should make this second transit across the gap. Thus the useful beam power was about

$$P_0 = (.53)\ (300)\ (.026) = 4.1.$$

If we assume a drift effectiveness factor F of unity, then for the $7\frac{3}{4}$ cycle mode, the efficiency should be given by H_m divided by $7\frac{3}{4}$. H_m is plotted as a function of G_R/y_e in Fig. 7. Thus, we can obtain η, the efficiency, and hence the power output. This has been done and the calculated power output is plotted vs W' in Fig. 54, where $W' = 1$ has been taken to correspond to 9,000 mc. It is seen that the theoretical variation of output with frequency is much the same as the measured variation.

Actually, of course, the parameters of the curve were chosen so that it corresponds fairly well to the experimental points. The upper value of W at which the tube goes out of oscillation is most strongly influenced by the value of θ_1 chosen. We see from Fig. 51 that as W is made greater than unity WD increases rapidly and hence, from Fig. 52, β^2 decreases rapidly, increasing G_R/y_e. On the other hand, as W is made smaller than unity, β^2 approaches unity but the grid loss term $W^{\frac{3}{2}}/D^2$ increases rapidly, and this term is most effective in adjusting the lower value of W at which oscillation will cease. Finally, the resonator loss term, varying as W^{-1}, does not change rapidly and can be used to adjust the total loss and hence the optimum value of G_R/y_e and the optimum efficiency.

It is clear that the power goes down at low frequencies chiefly because in moving the grids very close together to tune to low frequencies with a fixed nductance the resonator losses and especially the grid losses are increased.

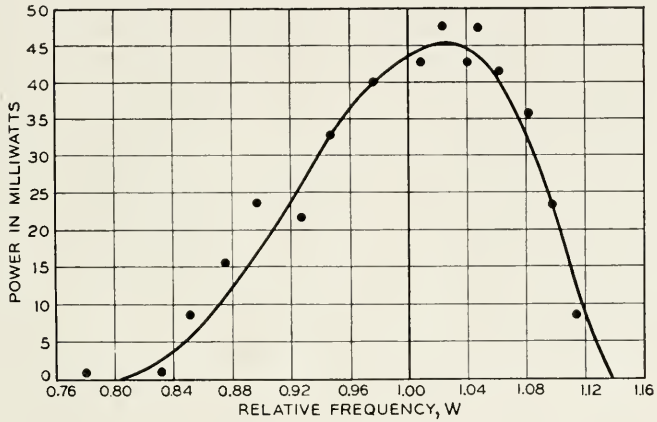


Fig. 54.—Computed curve of variation of power in milliwatts with relative frequency W for the parameters used in Fig. 53. The circles are experimental points. The curve has been fitted to the points by the choice of parameters.

In going to high frequencies the power decreases chiefly because moving the grids far apart to tune to high frequencies decreases β^2. Both of these effects are avoided if a fixed grid spacing is used and the tuning is accomplished by changing the inductance as in the case of the 707A. In such tubes there will be an upper frequency limit either because even with a fixed grid spacing β^2 decreases as frequency increases, or else there will be a limit at the resonant frequency of the smallest allowable external resonator, and there will be a lower frequency limit at which the repeller voltage for a given mode approaches zero; however, the total tuning range may be 3 to 1 instead of around 30% between extinction points, as for the 2K25.

The total electronic tuning between half-power points at optimum loading, $2(\Delta f)_{\frac{1}{2}}$, can be expressed

$$2(\Delta f)_{\frac{1}{2}} = (fy_e/M)(2\Delta\omega/\omega_0)/(y_e/M). \tag{10.15}$$

We can obtain $(2\Delta\omega/\omega_0)/(y_e/M)$ from Fig. 16.

If we assume a circuit consisting of a constant inductance L and a capacitance, the characteristic admittance of the resonator is

$$M = 1/\omega L = \tfrac{1}{2}\pi f_1 W \tag{10.16}$$

and

$$2(\Delta f)_{\frac{1}{2}} = 2\pi W^2 f_1^2 L y_e (2\Delta\omega/\omega_0)/(y_e/M) \tag{10.17}$$

and we have

$$y_e = \beta^2 I_0 (2\pi N)/2V_0 . \tag{10.18}$$

Here N is the total drift in cycles.

A rough calculation estimates the resonator inductance of the 2K25 as $.30 \times 10^{-9}$ henries. Using the values previously assumed, $I_0 = (.53)(.026)$, $V_0 = 300$, $N = 7\frac{3}{4}$, and the values of $G_R/y_e\beta^2$ and f_1 previously assumed, we can obtain electronic tuning.

A curve for half power electronic tuning vs W has been computed and is shown in Fig. 55, together with experimental data for a 2K25. The experimental data fall mostly above the computed curve. This could mean that the inductance has been incorrectly computed or that the drift effectiveness is increased over that for a linear drift field, possibly by the effects of space charge. By choosing a value of the drift effectiveness factor other than unity we could no doubt achieve a better fit of the electronic tuning data and still, by readjusting G_{g1} and G_{s1}, fit the power data. This whole procedure is open to serious question. Further, it is very hard to measure such factors as G_{g1} for a tube *under operating conditions*, with the grids heated by bombardment. Indirect measurements involve many parameters at once, and are suspect. Thus, Figs. 54 and 55 are presented merely to show a qualitative correspondence between theory and experiment.

XI. Noise Sidebands in Reflex Oscillations

In considering power production, the electron flow in reflex oscillators can be likened to a perfectly smooth flow of charge. However, the discrete nature of the electrons, the cause of the familiar "shot noise" in electron flow engenders the production of a small amount of *r-f* power in the neighborhood of the oscillating frequency—"noise sidebands". Thus the energy spectrum of a reflex oscillator consists of a very tall central spike, the power output of the oscillator, and, superposed, a distribution of noise energy having its highest value near the central spike.

Such noise or noise "sidebands" can be produced by any mechanism which causes the parameters of the oscillator to fluctuate with time. As the mean speed, the mean direction, and the convection current of the electron flow all fluctuate with time, possible mechanisms of noise production are numerous. Some of these mechanisms are:

(1) Fluctuation in mean speed causes fluctuation in the drift angle and hence can give rise to noise sidebands in the output through frequency modulation of the oscillator.

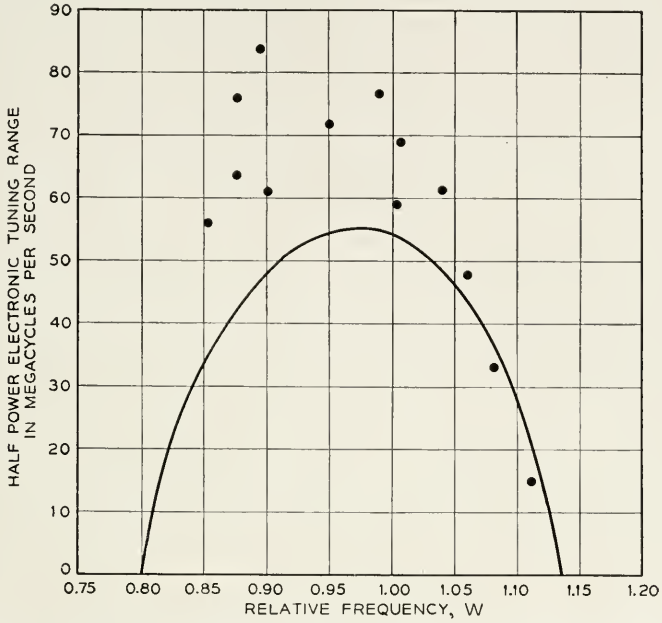


Fig. 55.—Computed variation of electronic tuning range in megacycles vs relative frequency W. The curve is calculated from the same data as that in Fig. 54 with no additional adjustment of parameters. Points represent experimental data.

(2) If the drift field acts differently on electrons differently directed, fluctuations in mean direction of the electron flow may cause noise sidebands through either amplitude or frequency modulation of the output.

(3) Low frequency fluctuations in the electron convection current may amplitude modulate the output, causing noise sidebands, and may frequency modulate the output when the oscillator is electronically tuned away from the optimum power point.

(4) High frequency fluctuations in the electron stream may induce high frequency noise currents in the resonator directly.

Mechanism (4) above, the direct induction of noise currents in the resonator by noise fluctuations in the electron stream, is probably most impor-

tant, although (3) may be appreciable. An analysis of the induction of noise in the resonator is surprisingly complicated, for the electron stream acts as a non-linear load impedance to the noise power giving rise to a complicated variation of noise with frequency and with amplitude of oscillation. On the basis of analysis and experience it is possible, however, to draw several general conclusions concerning reflex oscillator noise.

First, it is wise to decide just what shall be the measure of noise. The noise is important only when the oscillator is used as a beating oscillator, usually in connection with a crystal mixer. A power P is supplied to the mixer at the beating oscillator frequency. Also, the oscillator supplies at signal frequency, separated from the beating oscillator frequency by the intermediate frequency, a noise power P_n proportional, over a small frequency range, to the band-width B. An adequate measurement of the noisiness of the oscillator is the ratio of P_n to the Johnson noise power, kTB. The general facts which can be stated about this ratio and some explanation of them follow:

(1) Electrons which cross the gap only once contribute to noise but not to power. Likewise, if there is a large spread in drift angle among various electron paths, some electrons may contribute to noise but not to power.

(2) The greater the separation between signal frequency and beating oscillator frequency (i.e., the greater the intermediate frequency) the less the noise.

(3) The greater the electronic tuning range, the greater the noise for a given separation between signal frequency and beating oscillator frequency. This is natural; the electronic tuning range is a measure of the relative magnitudes of the electronic admittance and the characteristic admittance of the circuit.

(4) The degree of loading affects the noise through affecting the bunching parameter X. The noise seems to be least for light loading.

(5) Aside from controlling the degree of loading, resonator losses do not affect the noise; it does not matter whether the unused power is dissipated inside or outside of the tube.

(6) When the tube is tuned electronically, the noise usually increases at frequencies both above and below the optimum power frequency, but the tube is noisier when electronically tuned to lower frequencies. At the optimum frequency, the phase of the pulse induced in the circuit when an electron returns across the gap lags the pulse induced on the first crossing by 270°. When the drift time is shortened so as to tune to a higher frequency, the angle of lag is decreased and the two pulses tend to cancel; in tuning electronically to lower frequencies the pulses become more nearly in phase.

An approximate theoretical treatment leads to the conclusion that aside from avoiding loss of electrons in reflection, or very wide spreads in transit

time for various electrons, (see (1) above) and aside from narrowing the electronic tuning range, which may be inadmissable, the only way to reduce the noise is to decrease the cathode current. This is usually inadmissable. Thus, it appears that nothing much can be done about the noise in reflex oscillators without sacrificing electronic tuning range.

The seriousness of beating oscillator noise from a given tube depends, of course, on the noise figure of the receiver without beating oscillator noise and on the intermediate frequency. Usually, beating oscillator noise is worse at higher frequencies, partly because higher frequency oscillators have greater electronic tuning (see (3) above). At a wavelength of around 1.25 cm, with a 60 mc I.F. amplifier, the beating oscillator noise may be sufficient so that were there no other noise at all the noise figure of the receiver would be around 12 db.

Beating oscillator noise may be eliminated by use of a sharply tuned filter between the beating oscillator and the crystal. This precludes use of electronic tuning. Beating oscillator noise may also be eliminated by use of a balanced mixer in which, for example, the signal is fed to two crystals in the same phase and the beating oscillator in opposite phases. If the I.F. output is derived so that the signal components from the two crystals add, the output due to beating oscillator noise at signal frequencies will cancel out. There is an increasing tendency for a number of reasons to use balanced mixers and thus beating oscillator noise has become of less concern.

XII. Build-up of Oscillation

In certain applications, reflex oscillators are pulsed. In many of these it is required that the *r-f* output appear quickly after the application of *d-c* power, and that the time of build-up be as nearly the same as possible for successive applications of power. In this connection it is important to study the mechanism of the build-up of oscillations.

In connection with build-up of oscillations, it is convenient to use complex frequencies. Impedances and admittances at complex frequencies are given by the same functions of frequency as those at real frequencies. Suppose, for instance, the radian frequency is

$$\omega = w - j\alpha \tag{12.1}$$

This means the oscillations are increasing in amplitude. The admittance of a conductance G at this frequency is

$$Y = G$$

The admittance of a capacitance C and the impedance of an inductance L are

$$Y = j\omega C = jwC + \alpha C \tag{12.2}$$

$$Z = j\omega L = jwL + \alpha L \tag{12.3}$$

In other words, to an increasing oscillation reactive elements have a "loss" component of admittance or impedance. This "loss" component corresponds not to dissipation but to the increasing storage of electric or magnetic energy in the reactive elements as the oscillation increases in amplitude.

The admittance curves plotted in Figs. 41–46 may be regarded as contours in the admittance plane for $\alpha = 0$. If such a contour is known either by calculation or experiment, and it is divided into equal frequency increments, a simple construction will give a neighboring curve for $\omega = w - j\Delta\alpha$ where $\Delta\alpha$ is a small constant. Suppose that the change in Y for a frequency $\Delta\omega_1$ is ΔY_1. Then for a change $-j\Delta\alpha$

$$\Delta Y = -j \frac{\Delta Y_1}{\Delta\omega_1} \Delta\alpha. \tag{12.4}$$

Thus, to construct from a constant amplitude admittance curve an admittance curve for an increasing oscillation, one takes a constant fraction of each admittance increment between constant frequency increment points (a constant fraction of each space between circles in Figs. 41–46), rotates it 90 degrees clockwise, and thus establishes a point on the new curve.

This construction holds equally well for any conformal representation of the admittance plane (for instance, for the reflection coefficient plane represented on the Smith chart).

The general appearance of these curves for increasing oscillations in terms of the curve for real frequency can be appreciated at once. The increasing amplitude curve will lie to the right of the real frequency curve where the latter is rising and to the left where the latter is falling. Thus the loops will be diminished or eliminated altogether for increasing amplitude oscillations, and the low conductance portions will move to the right, to regions of higher conductance. This is consistent with the idea that for an increasing oscillation a "loss" component is added to each reactance, thus degrading the "Q", increasing the conductance, and smoothing out the admittance curve.

The oscillation starts from a very small amplitude, presumably that due to shot noise of the electron stream. For an appreciable fraction of the build-up period the oscillation will remain so small that nonlinearities are unimportant. The exponential build-up during this period is determined by the electronic admittance for very small signals.

As an example, consider a case in which the electronic admittance for small signals is a pure conductance with a value of $-y_e$. Here the fact that that the quantity is negative is recognized by prefixing a minus sign.

Assume also that the circuit admittance including the load may be expressed as in (a-22) of Appendix I, which holds very nearly in case there is only one resonance in resonator and load. Then for a complex frequency $w_o - j\alpha_o$ the circuit admittance will be

$$Y_c = G_c + 2M\alpha_o/w_o \tag{12.5}$$

Thus in this special case we have for oscillation

$$y_{eo} = G_c + 2M\alpha_o/w_o \tag{12.6}$$

and

$$\alpha_o = \frac{w_o}{2M}(Y_{eo} - G_c). \tag{12.7}$$

The amplitude, then, builds up initially according to the law

$$V = V_o e^{\alpha_o t}. \tag{12.8}$$

If the amplitude does not change too rapidly, the build-up characteristic of an oscillator can be obtained step-by-step from a number of contours for constant α and from a $-Y_e$ curve marked with amplitude points. The $-Y_e$ curve might, for instance, be obtained from a Rieke diagram and an admittance curve.

Consider the example shown in Fig. 56. Fig. 56a shows curves constructed for complex frequencies from the admittance curve for the resonant circuit for real frequency. In addition the negative of the electronic admittance is shown. Oscillation will start from some very small amplitude, $V = V_o$, and build-up at an average rate given by $\alpha = 2.5 \times 10^{-6}$ until $V = 1$. Let $V_o = .1$. Then the interval to build-up from $V = .1$ to $V = 1$ is

$$\Delta t_1 = \frac{\ln\left(\frac{1}{.1}\right)}{2.5 \times 10^{-6}}$$

$$= .92 \times 10^{-6} \text{ seconds.}$$

From amplitude 1 to amplitude 2 the average value of α will be 1.5×10^{-6} and the time interval will be

$$\Delta t_2 = \frac{\ln\left(\frac{2}{1}\right)}{1.5 \times 10^{-6}}$$

$$= .46 \times 10^{-6} \text{ seconds.}$$

Similarly, from 2 to 3

$$\Delta t_3 = \frac{\ln\left(\frac{3}{2}\right)}{.5 \times 10^{-6}}$$

$$= .80 \times 10^{-6} \text{ seconds.}$$

The build-up curve is shown in Fig. 56b.

Similarly, from a family of admittance contours constructed from a cold impedance curve, and from a knowledge of frequency and amplitude vs time,

Y_e can be obtained as a function of time. It may be that in many cases the real part of the frequency is nearly enough constant during build-up so that only the amplitude vs time need be known. As the input will commonly be a function of time for such experimental data, Y_e vs time will yield Y_e at vari-

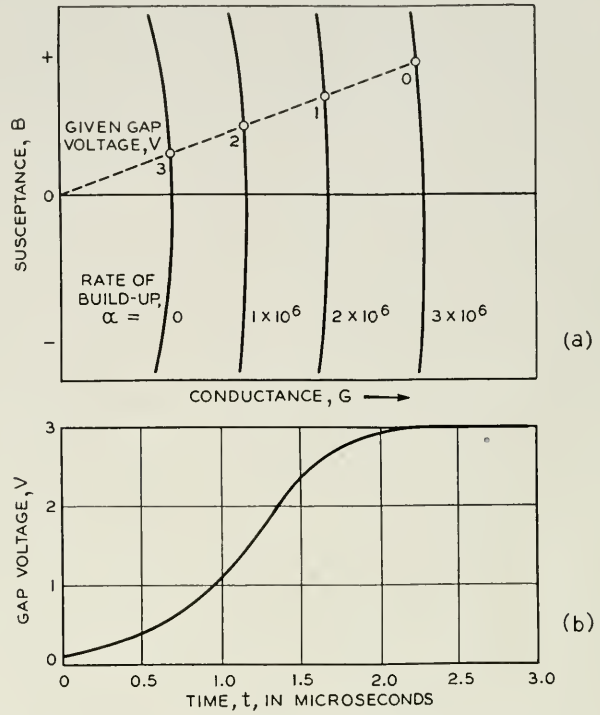


Fig. 56.—*a*. A plot of the circuit admittance (solid lines) for various rates of build-up specified by the parameters α. The voltage builds up as $e^{\alpha t}$. The circuit conductance is greater for large values of α. The negative of the electronic admittance is shown by the dashed lines. The circles mark off the admittance at which various amplitudes or voltages of oscillation occur. The intersections give the rates of build-up of oscillation at various voltages. By assuming exponential build up at a rate specified by α between the voltages at these intersections, an approximate build-up can be constructed.

b. A build up curve constructed from the data in Fig. 56a.

ous amplitudes and inputs. Curves for various rates of applying input will yield tables of Y_e as a function of both input and amplitude.

It will be noted that to obtain very fast build-up with a given electronic admittance, the conductance should vary slowly with α. This is the same as saying that the susceptance should vary slowly with ω, or with real frequency. For singly resonant circuits, this means that ω_o/M should be large.

Suppose the admittance curve for real frequency, i.e. $\alpha = 0$, has a single

loop and is symmetrical about the G axis as shown in Fig. 57. Suppose the $-V_e$ curve lies directly on the G axis. The admittance contours for increasing values of α will look somewhat as shown. Suppose build-up starts on Curve 2. When Curve 1 with the cusp is reached, the build-up can continue along either half as the loop is formed and expands, resulting either of the two possible frequencies of Curve 0. Presumably in this symmetrical

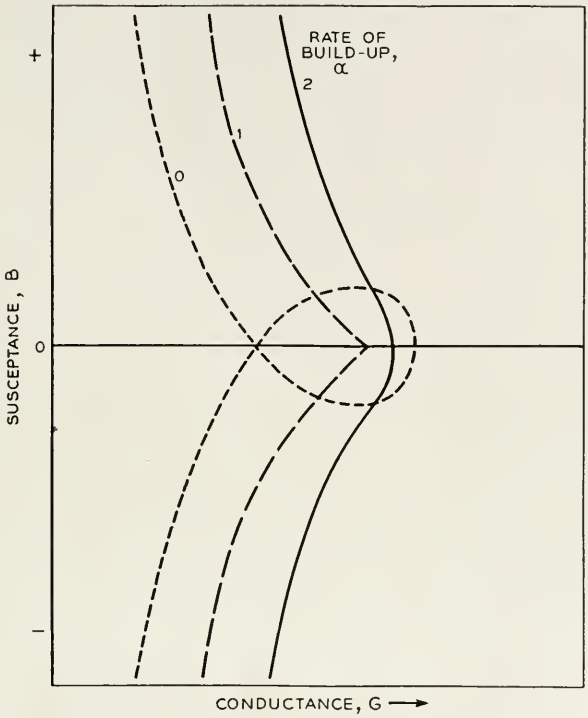


Fig. 57.—Circuit admittance vs circuit conductance in arbitrary units for different rates of build-up at turn-on. When the build-up is rapid ($\alpha = 2$) the admittance curve has no loop. As the rate of build-up decreases the curve sharpens until it has a cusp $\alpha = 1$. As the rate of build-up further decreases the curve develops a loop ($\alpha = 0$). There may be uncertainty as to which of the final intersections with the $\alpha = 0$ line will represent oscillation.

case, nonsynchronous fluctuations would result in build-up to each frequency for half of the turn-ons. If one frequency were favored by a slight dissymmetry, the favored frequency would appear on the greater fraction of turn-ons. For a great dissymetry, build-up may always be in one mode, although from the impedance diagram steady oscillation in another mode appears to be possible.

In the absence of hum or other disturbances the build-up of oscillations starts from a randomly fluctuating voltage caused by shot noise.[11] Thus, from turn-on to turn-on some sort of statistical distribution may be expected in the time τ taken to reach a given fraction of the final amplitude. In unpublished work Dr. C. R. Shannon of these laboratories has shown that in terms of α_o, the initial rate of build-up, the standard deviation $\delta\tau$ and the root mean square deviation $(\overline{\delta\tau^2})^{1/2}$ are given by

$$\delta\tau = .38/\alpha_o \tag{12.9}$$

$$(\overline{\delta\tau^2})^{1/2} = .64/\alpha_o \tag{12.10}$$

Thus the "jitter" in the successive positions of the r-f pulses associated with evenly spaced turn-ons is least when the initial rate of build-up, given by α_o, is greatest.

Such conditions do not obtain on turn-off, and there is little jitter in the trailing edge of a series of r-f pulses. This is of considerable practical importance.

XIII. Reflex Oscillator Development at the Bell Telephone Laboratories

For many years research and development directed towards the generation of power at higher and higher frequencies have been conducted at the Bell Telephone Laboratories. An effort has been made to extend the frequency range of the conventional grid controlled vacuum tube as well as to explore new principles, such as those embodied in velocity variation oscillators. The need for centimeter range oscillators for radar applications provided an added impetus to this program and even before the United States entry into the war, as well as throughout its duration, these laboratories, cooperating with government agencies, engaged in a major effort to provide such power sources. The part of this program which dealt with high power sources for transmitter uses has been described elsewhere. This paper deals with low power sources, which are used as beating oscillators in radar receivers. In the following sections some of the requirements on a beating oscillator for a radar receiver will be outlined in order to show how the reflex oscillator is particularly well suited for such an application.

A. *The Beating Oscillator Problem*

The need for a beating oscillator in a radar system arises from the necessity of amplifying the very weak signals reflected from the targets. Immediate rectification of these signals would entail a very large degradation in signal to noise ratio, although providing great simplicity of operation. It would also lead to a lack of selectivity. Amplification of the signals at the

[11] See Appendix 10.

signal frequency would require centimeter range amplifiers having good signal to noise properties. No such amplifiers existed for the centimeter range, and it was necessary to beat the signal frequency to an intermediate frequency for amplification before rectification. For a number of reasons, such intermediate frequency amplifiers operate in the range of a few tens of megacycles, so that the beating oscillator must generate very nearly the same frequency as the transmitter oscillator.

In radar receivers operating at frequencies up to several hundred megacycles, conversion is frequently achieved with vacuum tubes. For higher frequencies crystal converters have usually been employed. With few exceptions, the oscillators to be described were used with these crystal converters which require a small oscillator drive of the order of one milliwatt. In general it is desirable to introduce attenuation between the oscillator and the crystal to minimize effects due to variation of the load. Approximately 13 db is allowed for such padding so that a beating oscillator need supply about 20 milliwatts. Power in excess of this is useful in many applications but not absolutely necessary. Since the power output requirements are low, efficiency is not of prime importance and is usually, and frequently necessarily, sacrificed in the interest of more important characteristics.

The beating oscillator of a radar receiver operating in the centimeter range must fulfill a number of requirements which arise from the particular nature of the radar components and their manner of operation. The intermediate frequency amplifier must have a minimum pass band sufficient to amplify enough of the transmitter sideband frequencies so that the modulating pulse is reproduced satisfactorily. It is not desirable to provide much margin in band width above this minimum since the total noise increases with increasing band width. It is therefore necessary for best operation that the frequency of the beating oscillator should closely follow frequency variations of the transmitter, so that a constant difference frequency equal to the intermediate frequency is maintained.

This becomes more difficult at higher frequencies, inasmuch as all frequency instabilities, such as thermal drifts, frequency pulling, etc. occur as percentage variations. Some of the frequency variations occur at rapid rates. An example of this is the frequency variation which is caused by changes in the standing wave presented to the transmitter. Such variations may arise, for instance, from imperfections in rotating joints in the output line between the transmitter to the scanning antenna.

For correction of slow frequency drifts a manual adjustment of the frequency is frequently possible, but instances arise, notably in aircraft installations, in which it is not possible for an operator to monitor the frequency constantly. Rapid frequency changes, moreover, occur at rates in excess of the reaction speed of a normal man. Hence for obvious tactical reasons

it is imperative that the difference frequency between the transmitter and the beating oscillator should be maintained by automatic means. As an illustration of the problem one may expect to have to correct frequency shifts from all causes, in a 10,000 megacycle system, of the order of 20 megacycles. Such correction may be demanded at rates of the order of 100 megacycles per second per second.

Although the frequency range of triode oscillators has since been somewhat extended, at the time that beating oscillators in the 10 centimeter range were first required the triode oscillators available did not adequately fulfill all the requirements. In general the tuning and feedback adjustments were complicated and hence did not adapt themselves to automatic frequency control systems. Velocity variation tubes of the multiple gap type which gave more satisfactory performance than the triodes existed in this range. These, however, generally required operating voltages of the order of a thousand volts and frequently required magnetic fields for focussing the electron stream. The tuning range obtainable by electrical means was considerably less than needed and, just as in the case of the triode oscillator, the mechanical tuning mechanism did not adapt itself to automatic control. These difficulties focussed attention on the reflex oscillator, whose properties are ideally suited to automatic frequency control. The feature of a single resonant circuit is of considerable importance in a military application, in which simple adjustments are of primary concern. The repeller control of the phase of the negative electronic admittance which causes oscillation provides a highly desirable vernier adjustment of the frequency, and, since this control dissipates no power, it is particularly suited to automatic frequency control. Furthermore, since the upper limit on the rate of change of frequency is set by the time of transit of the electrons in the repeller field and the time constant of the resonant circuit, both of which are generally very small fractions of a micro-second, very rapid frequency correction is possible.

As the frequency is varied with the repeller voltage, the amplitude of oscillation also varies in a manner previously described. The signal to noise performance of a crystal mixer depends in part on the beating oscillator level and has an optimum value with respect to this parameter. In consequence, there are limitations on how much the beating oscillator power may depart from this optimum value. This has a bearing on the oscillator design in that the amount of amplitude variation permitted for a given frequency shift is limited. The usual criterion of perfomance adopted has been the electronic tuning, i.e. the frequency difference, between points for a given repeller mode at which the power has been reduced to half the maximum value.

Reception of the wrong sideband by the receiver causes trouble in con-

nection with automatic frequency control circuits in a manner too complicated for treatment here. In some cases this necessitates a restriction on the total frequency shift between extinction points for a given repeller mode. The relationship between half power and extinction electronic tuning has been discussed in Section VII.

In addition to the electrical requirements which have been outlined, military applications dictate two further major objectives. The first is the attainment of simple installation and replacement, which will determine, in part, the outward form of the oscillator. The second is low voltage operation, which fundamentally affects the internal design of the tubes. In some instances military requirements conflict with optimum electronic and circuit design, and best performance had to be sacrificed for simplicity of construction and operation. In particular, in some cases it was necessary to design for maximum flexibility of use and compromise to a certain extent the specific requirements of a particular need.

In the following section we will describe a number of reflex oscillators which were designed at the Bell Telephone Laboratories primarily to meet military requirements. These oscillators are described in approximate chronological order of development in order to indicate advances in design and the factors which led to these advances.

The reflex oscillators which will be described fall into two general classifications determined by the method employed in tuning the resonator. In one category are oscillators tuned by varying primarily the inductance of the resonator and in the other are those tuned by varying primarily the capacitance of the resonator. The second category includes two types in which the capacitance is varied in one case by external mechanical means and in the second case by an internal means using a thermal control.

B. *A Reflex Oscillator With An External Resonator—The 707*

The Western Electric 707A tube, which was the first reflex oscillator extensively used in radar applications, is characteristic of reflex oscillators using inductance tuning. It was intended specifically for service in radar systems operating at frequencies in a range around 3000 megacycles. Fig. 58 shows a photograph of the tube and Fig. 59 an x-ray view showing the internal construction. A removable external cavity is employed with the 707A as indicated by the sketch superimposed on the x-ray of Fig. 59. Such cavities are tuned by variation of the size of the resonant chamber. Such tuning can be considered to result from variation of the inductance of the circuit.

The form of this oscillator is essentially that of the idealized oscillator shown in Fig. 58. The electron gun is designed to produce a rectilinear cylindrical beam. The gun consists of a disc cathode, a beam forming elec-

Fig. 58.—External view of the W.E. 707-A reflex oscillator tube. This tube is intended for use with an external cavity and was the first of a series of low voltage oscillators.

trode and an accelerating electrode G_1 which is a mesh grid formed on a radius. The gun design is based on the principle of maintaining boundary conditions such that a rectilinear electron beam will flow through the resonator gap. The resonator grids G_2 and G_3 are mounted on copper discs.

One of these has a re-entrant shape to minimize stray capacitance in the resonant circuit. These discs are sealed to glass tubing which provides a

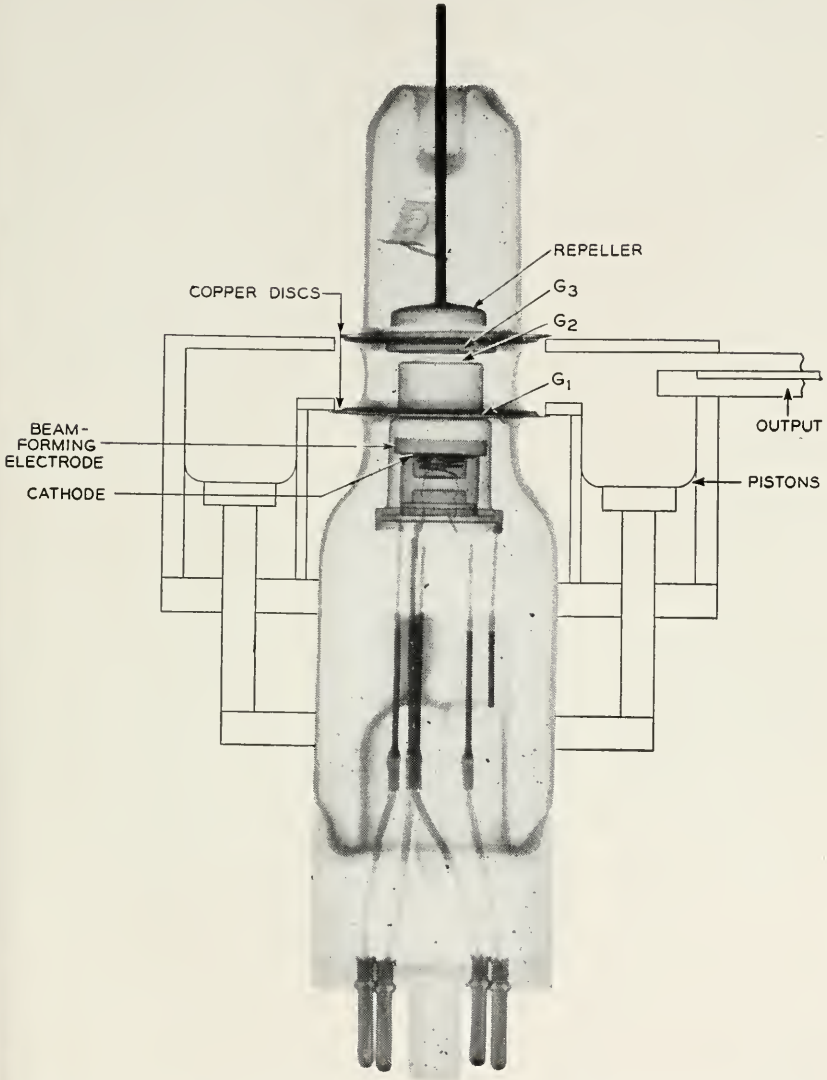


Fig. 59.—X-ray view of the W.E. 707-A shows the method of applying an external cavity tuned with a piston.

vacuum envelope. The discs extend beyond the glass to permit attachment to the external resonant chamber. The shape of the repeller is chosen

to provide as nearly as possible a uniform field in the region into which the beam penetrates.

A wide variety of cavity resonators has been designed for use with this oscillator. An oscillator of this construction is fundamentally capable of oscillating over a much wider frequency range than tubes tunable by means of capacitance variation. The advantage arises from the fact that the interaction gap where the electron stream is modulated by the radio frequency field is fixed. As discussed in more detail in Section X, this results in a slower variation of the modulation coefficient with frequency and also a slower variation of cavity losses and gap impedance than in an oscillator in which tuning is accomplished by changing the gap spacing. A cavity designed for wide range frequency coverage using the 707A tube is shown in Fig. 60. Using such a cavity it is possible to cover a frequency range from 1150 to 3750 megacycles. The inductance of the circuit is varied by moving the shorting piston in the coaxial line. For narrow frequency ranges,

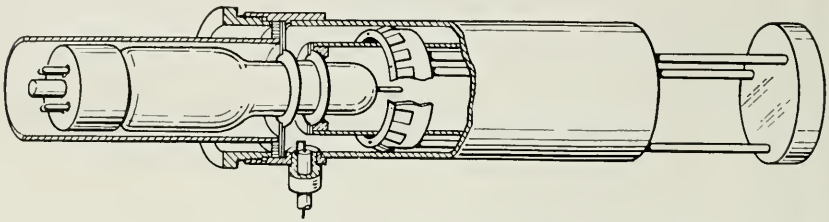


Fig. 60.—Sketch showing a piston tuned circuit for the W.E. 707-A which will permit operation from 1150 to 3750 mc.

cavities of the type shown in Fig. 61 are more suitable. In such cavities tuning is effected by means of plugs which screw into the cavity to change its effective inductance. Power may be extracted from the cavity by means of an adjustable coupling loop as shown in Fig. 61.

The 707A was the first reflex oscillator designed to operate at a low voltage i.e. 300 volts. This low operating voltage proved to be a considerable advantage in radar receivers because power supplies in this voltage range provided for the i.f. amplifiers could be used for the beating oscillator as well. Operation at this voltage was achieved by using an interaction gap with fine grids, which limits the penetration of high frequency fields. This results in a shorter effective transit angle across the gap for a given gap spacing and a given gap voltage than for a gap with coarse or no grids. Hence, for a given gap spacing a good modulation coefficient can be obtained at a lower voltage. Moreover, since drift action results in more efficient bunching at low voltages, a larger electronic admittance is obtained than with an open gap. This gain in admittance more than outweighs the

greater capacitance of a gap with fine grids, so that a larger electronic tuning range is obtained than with an open gap. The successful low voltage operation of the 707A established a precedent which was followed in all the succeeding reflex oscillators designed for radar purposes at the Bell Tele-

Fig. 61.—A narrow tuning range cavity for the W.E. 707-A of the type used in radar systems. The inductance of the cavity can be adjusted by moving screws into it. This view also shows the adjustable coupling loop.

phone Laboratories. The 707A is required to provide a minimum power output of 25 milliwatts and a half power electronic tuning of 20 megacycles near 3700 megacycles. The power output and the electronic tuning are in excess of this value over the range from 2500 megacycles to 3700 megacycles in a repeller mode having $3\frac{3}{4}$ cycles of drift.

C. *A Reflex Oscillator With An Integral Cavity—The* 723

The need for higher definition in radar systems constantly urges operation at shorter wavelengths. Thus, while radar development proceeded at 3000 megacycles, a program of development in the neighborhood of 10,000 megacycles was undertaken. Although waveguide circuit techniques were employed to some extent at 3000 megacycles, the cumbersome size of the guide made its use impractical in the receiver and hence coaxial techniques were employed. The 1″ by $\frac{1}{2}$″ guide used at 10,000 megacycles is convenient in receiver design and also desirable because the loss in coaxial conductors becomes excessive at this frequency. Hence, one of the first requirements on an oscillator for frequencies in this range was the adaptability of the output circuit to waveguide coupling.

In considering possible designs for a 10,000 megacycle oscillator the simple scaling of the 707A was studied. This appeared impractical for a number of reasons. The most important limitation was the constructional difficulty of maintaining the spacing in the gap with sufficient accuracy with the glass sealing technique available. Also, variations in the capacitance caused by variations in the thickness of the seals caused serious difficulties in predetermining an external resonator. Contributing difficulties arose from the power losses in the glass within the resonant circuit and the problem of making the copper to glass seals close to the internal elements.

Consideration of these factors led to a new approach to the problem, in which the whole of the resonant circuit was enclosed within the vacuum envelope. This required a different mechanism for tuning the resonator, since variation of the inductance of a cavity requires relatively large displacements which are difficult to achieve through vacuum seals. The alternative is to vary the capacitance of the gap. Since the gap is small a relatively large change in capacitance can be achieved with a small displacement. This sort of tuning permits the use of metal tube constructional techniques, and these were applied.

As a matter of historical interest an attempt at this technique made at the Bell Telephone Laboratories is shown in Fig. 62. This device was held together by a sealing wax and string technique and was not tunable in the first version. It oscillated successfully on the pumps, however, and a second version was constructed which was tuned by means of an adjustable coaxial line shunting the cavity resonator. Adjustment of this auxiliary line gave a tuning range of 7.5%. Such a tuning method is fraught with the complications outlined in Section IX.

An early reflex oscillator tube of the integral cavity type designed at the Bell Telephone Laboratories was the Western Electric 723A/B.

This design was superseded later by the W.E. 2K25 which has a greater frequency range and a number of design refinements. From a constructional point of view the two types are closely similar, however, and to avoid duplication the later tube will be described to typify a construction which served as a basis for a whole series of oscillators in the range from 2500 to 10,000 Mc/s.

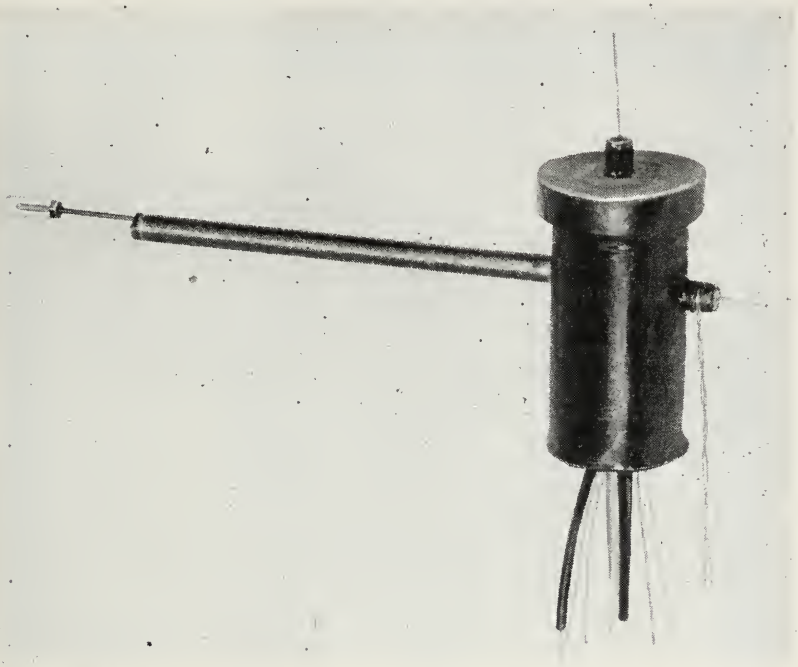


Fig. 62.—An early continuously pumped metal reflex oscillator tuned with an external line.

Before proceeding to a description of the 2K25 tube it seems desirable to recapitulate in more detail the design objectives from a mechanical point of view. These were:

1. To provide a design which would lend itself to large scale production and one sufficiently rugged as to be capable of withstanding the rough use inherent in military service.
2. To provide output means which permit coupling to a wave guide in such a manner that installation or replacement could be accomplished in the simplest possible manner.
3. To provide a tuning mechanism for the resonant circuit which, while simple, would give sufficiently fine tuning to permit setting and holding

a frequency within one or two parts in 10,000. In addition, in order to avoid field installation it was desired to have the tuning mechanism cheap enough to be factory installed and discarded with each tube.

4. The oscillator was required to be compact and light in weight to facilitate its use in airborne and pack systems.

Figure 63 shows a cross-section view of the final design of the 2K25 reflex oscillator. The resonant cavity is formed in part by the volume included between the frames which support the cavity grids and also by the volume between the flexible vacuum diaphragm and one of the frames. This diaphragm also supports a vacuum housing containing the repeller. The electron optical system consists of a disc cathode, a beam electrode and an accelerating grid. These are so designed as to produce a slightly convergent outgoing electron stream. The purpose of this initial convergence is to offset the divergence of the stream caused by space charge after the stream passes the accelerating grid and to minimize the fraction of the electron stream captured on the grid frame on the round trip. The repeller is designed to provide as nearly as possible a uniform retarding field through the stream cross-section.

Power is extracted from the resonant circuit by the coupling loop and is carried by the coaxial line to the external circuit. The center conductor of the coaxial line external to the vacuum is supported by a polystyrene insulator and extends beyond the outer conductor to form a probe. Coupling to a wave guide is accomplished by projecting this probe through a hole in that wall of the wave guide which is perpendicular to the E lines so that the full length of the probe extends into the guide. The outer conductor is connected to the wave guide either metallically or by means of an r.f. bypass or choke circuit. A more detailed section on such coupling methods will be given later.

The tube employed a standard octal base modified to pass the coaxial line. Thus if a standard octal socket is similarly modified and mounted on the wave guide it is possible to couple the oscillator to the wave guide and power supply circuits simply by plugging it into the socket, just as with any conventional vacuum tube.

The tuning means for this type of oscillator tube presented a serious problem. This will be appreciated when it is realized that the mechanism must permit setting frequencies correctly to within one megacycle in a device in which the frequency changes at the rate of approximately 200 megacycles per thousandth of an inch displacement of the grids. In other words, the tuner was required to make possible the adjustment of the grid spacing to an accuracy of five millionths of an inch. The design of the mechanism adopted was originated by Mr. R. L. Vance of these Laboratories. The operation of the tuner can be seen from an examination of the cross-section

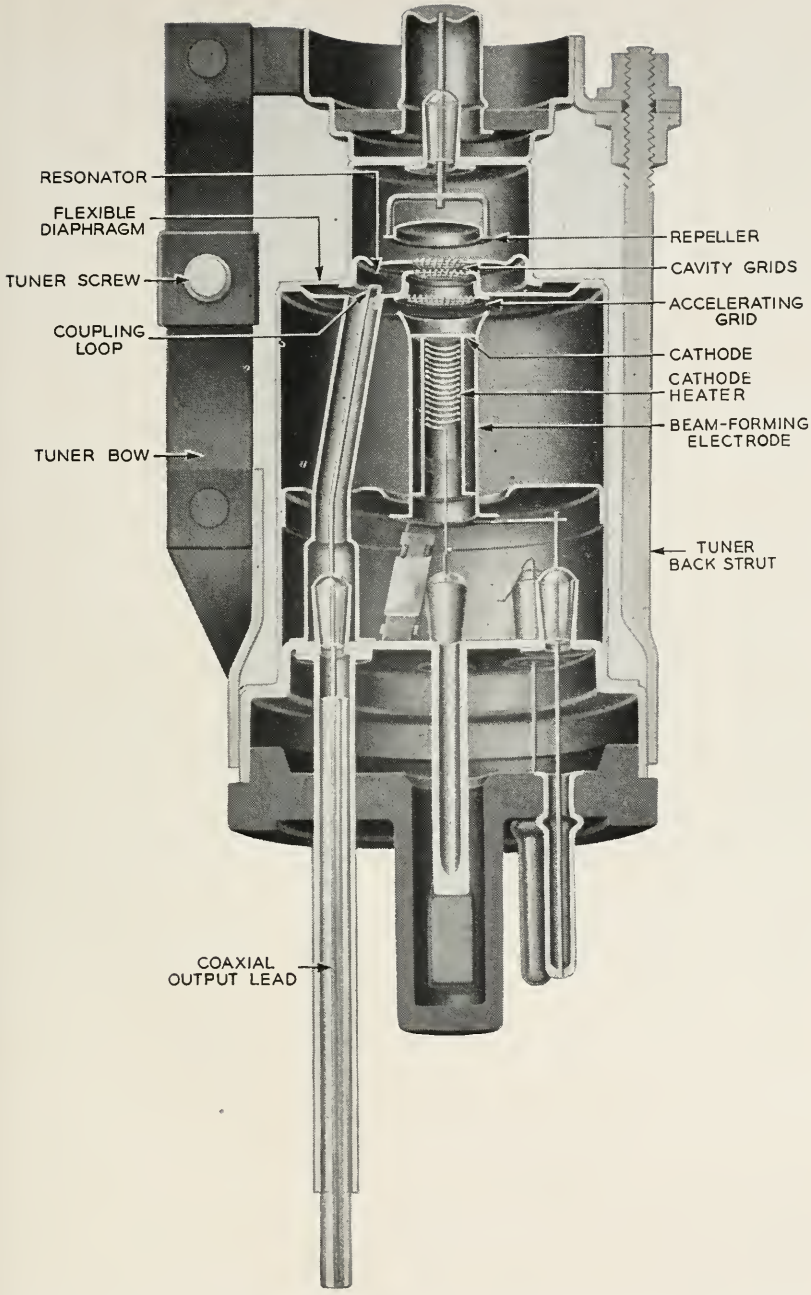


Fig. 63.—A 3 centimeter reflex oscillator with an internal resonator. This tube is a further development of the earliest internal resonator reflex oscillator designed at the Bell Telephone Laboratories.

and external views of Figs. 63 and 67. On one side of the tube a strut extending from the base is attached to the repeller housing. This strut acts as a rigid vertical support but provides a hinge for lateral motion. On the opposite side the support is provided by a pair of steel strips. These are clamped together where they are attached to the vacuum housing support and also where they are attached to a short fixed strut near the base. A nut is attached rigidly to the center of each strip. One nut has a right and the other a left handed thread. A screw threaded right handed on one half and left handed on the other half of its length turns in these nuts and drives them apart. The mechanism is thus a toggle which, through the linkage provided by the repeller housing, serves to move the grids relative to one another and thus to provide tuning action.

The 723A/B was originally designed for a relatively narrow band in the vicinity of 9375 megacycles. It operates at a resonator voltage of 300 volts and the beam current of a typical tube would be approximately 24 milliamperes. The design was based on the use of repeller voltage mode which with the manufacturing tolerances lay between 130 and 185 volts at 9375 megacycles. It is difficult to establish with certainty the number of cycles of drift for this mode. Experimental data can be fitted by values of either $6\frac{3}{4}$ or $7\frac{3}{4}$ cycles and various uncertainties make the value calculated from dimensions and observed voltages equally unreliable. This value is, however, of interest principally to the designer and of no particular moment in application. The performance was specified for the output line of the oscillator coupled to a $\frac{5}{8}'' \times 1\frac{1}{4}''$ wave guide so that the probe projected full length into the guide through the wider wall and on the axis of the guide. With a matched load coupled in one direction and a shorting piston adjusted for an optimum in the other the oscillator was required to deliver a minimum of 20 milliwatts power output at a frequency of 9375 megacycles. Under the same conditions the electronic tuning was required to be at least 28 megacycles between half power points.

For reasons of continuity a more detailed description of the properties of the 3 centimeter oscillator will be given in a later section. The 723A/B oscillator served as the beating oscillator for all radar systems operating in the 3 centimeter range until late in the war when the 2K25 supplanted it. At the time that the 723A/B was developed the best techniques and equipment available were employed. In retrospect these were somewhat primitive and of course this resulted in a number of limitations of performance. Since the tubes designed as beating oscillators commonly served as signal generators in the development of ultra-high frequency techniques and equipment the wartime designer of such oscillators usually found himself in the position of lifting himself by his own bootstraps. In spite of these limitations the later modifications of the 723A/B which led to

the 2K25 did not fundamentally change the design but were rather in the direction of extending its performance to meet the expanding requirements of the radar art. The incorporation of the resonant cavity within the vacuum envelope resulted in a major revision of the scope of the designer's problems. He assumed a part of the burden of the circuit engineer in that it became necessary for him to design an appropriate cavity and predetermine the correct coupling of the oscillator to the load. The latter transferred to the laboratory a problem which in the case of separate cavity oscillators had been left as a field adjustment.

D. *A Reflex Oscillator Designed to Eliminate Hysteresis—The 2K29*

As service experience with the external cavity type of reflex oscillator was gained a number of limitations of such a design became apparent. The difficulties arose primarily from the conditions of military application. A typical difficulty was the corrosion of cavities and copper flanges under the severe tropical conditions met in some service applications. The difficulty of maintaining a moistureproof seal in a cavity tuned by variation of the inductance made it very difficult to alleviate this condition. The success of the all metal technique in the three centimeter range suggested the application of the same principles to the design of a 10 centimeter oscillator and this was undertaken.

Mechanically, the problem was straightforward, but an extrapolation of the electrical design of the 723A/B to 10 centimeters suffered from a fatal defect. The difficulty, previously described in Section VIII, was the discontinuous and multiple valued character of the output as a function of the repeller voltage. Reference to Fig. 19 will indicate the operational problems which would arise in an oscillator in which the hysteresis existed in marked degree. The a.f.c. systems were such that in starting the repeller voltage would start from a value more negative than required for oscillation and decrease. As the repeller voltage decreased through the range where oscillation would occur the frequency would of course cover a range of values. When the repeller voltage reached a value such that the frequency of the oscillator had a value differing from the transmitter by the intermediate frequency the steady shift of the repeller voltage would be stopped and would then hunt over a limited range about the value required to maintain the difference frequency. When adjusted for operation this condition would pertain with the repeller voltage at a value such that the oscillator would be delivering maximum power. If under operating conditions the frequency required of the oscillator by the system drifted to that corresponding to the amplitude jump at B, any further drift of frequency could not be corrected. Thus, one effect of the hysteresis is to limit the electronic tuning range. As a second possibility, let us assume that the

frequency has drifted so that the oscillator is operating in a range between A and B of Fig. 19. If now the operation of the system is momentarily interrupted the a.f.c. system will start hunting. This is done by returning to the non-oscillating repeller voltage just as when operation is initiated. When the hunting repeller voltage passes through the value between A and B from the non-oscillating state no oscillation occurs and hence the a.f.c. cannot lock in and the system becomes inoperative. Thus it is imperative that hysteresis be kept, as a minimum requirement, outside the useful electronic tuning range.

As indicated in Section VIII it was found that the electronic hysteresis occurred when the electron stream made more than two transits across the gap. Thus an added objective of the design of the 10 centimeter metal oscillator became the achievement of an electron optical system which would limit the number of transits to two while insuring that the maximum number of electrons leaving the cathode would make the two transits with a minimum spread in zero signal transit time.

Figure 64 shows a sectional view of the final design adopted. The electron optical structure differs from that of the 723A/B in a number of respects. The first grid of the 723A/B has been eliminated and one of the cavity grids now plays a dual role in simultaneously serving as an accelerating grid. The grids are curved towards the cathode, which has a central spike. This arrangement is intended to produce a hollow cylindrical electron stream. It will be observed that the second grid is larger in diameter than the first and that the repeller has a central spike. The design is such that the cylindrical beam entering the repeller region is caused to diverge radially, so that in re-traversing the gap after its reversal in direction it impinges on and is captured by the frame supporting the first grid.

The repeller design was determined by using an electrolytic trough to determine the potential plots for a number of trial configurations. Then by making point by point calculations of the electron paths the best configuration was chosen. A typical example of such path tracing is shown in Fig. 65. This figure shows the equipotential lines and the trajectories computed for electrons on the inner and outer boundaries of the outgoing stream. The method of calculating the trajectories has been described by Zworykin and Rajchman.[12] It assumes that space charge may be neglected. Fig. 65 shows that the repeller design is such that the cylindrical outgoing stream is focussed upon its return onto the frame supporting the first grid. The cathode spike prevents emission from the central portion of the cathode, since it would be difficult to prevent electrons from this portion from returning into the cathode space. A second requirement on

[12] V. K. Zworykin and J. A. Rajchman, *Proc. of I.R.E.*, Sept. 1939, Vol. 17, No. 9, pp. 558–566.

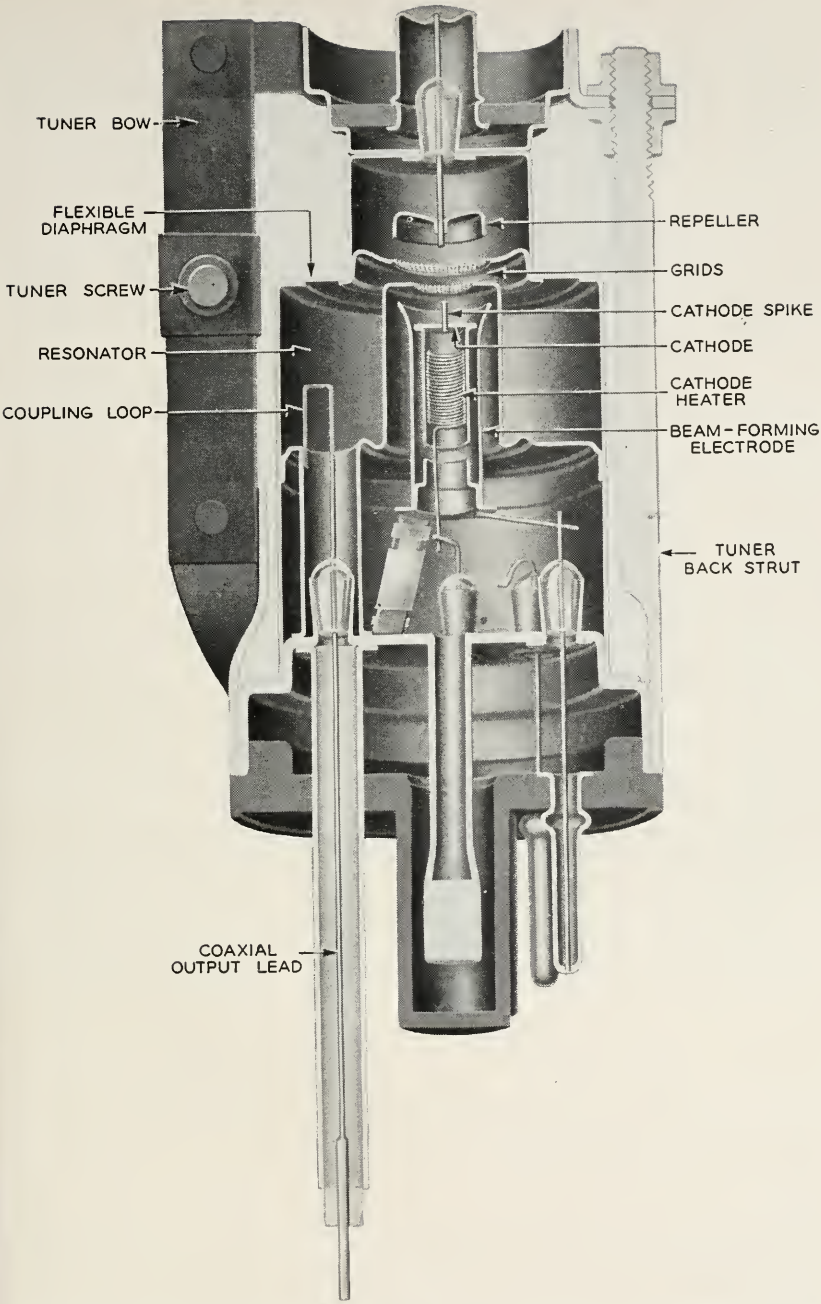


Fig. 64.—Section view of the W.E. 2K29 reflex oscillator shows the electron optical system used in eliminating hysteresis. (Fig. 65) This tube has an internal cavity and is designed for the frequency range from 3400 to 3960 mc/s.

the design was that the spread in the transit angle for zero signal should be small. This requirement is not as stringent as might be expected. The contribution to the electronic conductance by a current element whose

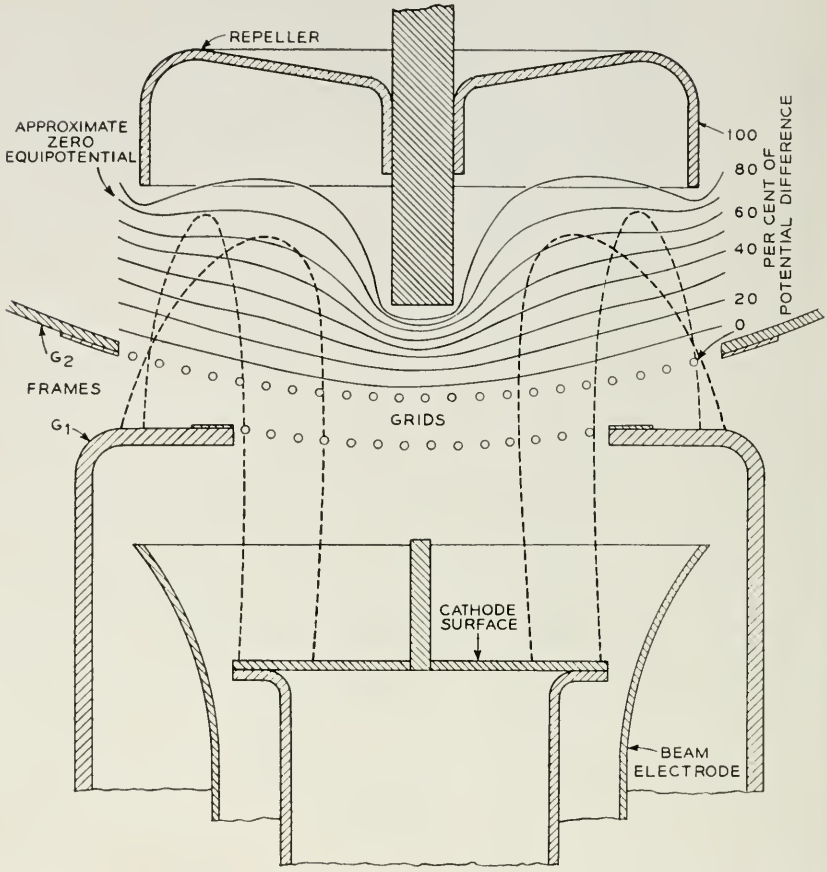


Fig. 65.—Repeller design for eliminating hysteresis in a reflex oscillator. The electrons make only two transits through the gap. The repeller does not return them to the cathode region but to the edge supporting one of the grids of the gap. Equipotentials determined from an electrolytic trough investigation are shown and the electron trajectories computed from these equipotentials.

transit angle deviates from the optimum by a small angle $\Delta\theta$ varies as $\cos \Delta\theta$, so that even for spreads in angle as great as $\pm$ 30° the effect is not serious.

The design illustrated in Fig. 64 was strikingly effective in reducing hysteresis. Fig. 66 shows repeller characteristics for the original design which was extrapolated from the 723A/B and for the design of Fig. 64.

In addition to improving the electronic tuning characteristics the design was found to be more stable against variations in load, as would be expected from the discussion of Section VIII.

The arrangement adcpted provided a prototype electron optical system which was used in a whole series cf reflex oscillators designed for radar and communication systems. These tubes were the 726A, 726B and 726C, 2K29, 2K22, 2K23 and 2K56.

The output line of these tubes is intended to couple through an adapter to either a coaxial line or a wave guide. In the first case the adapter serves to couple the output line of the tube to the cable, in some instances through

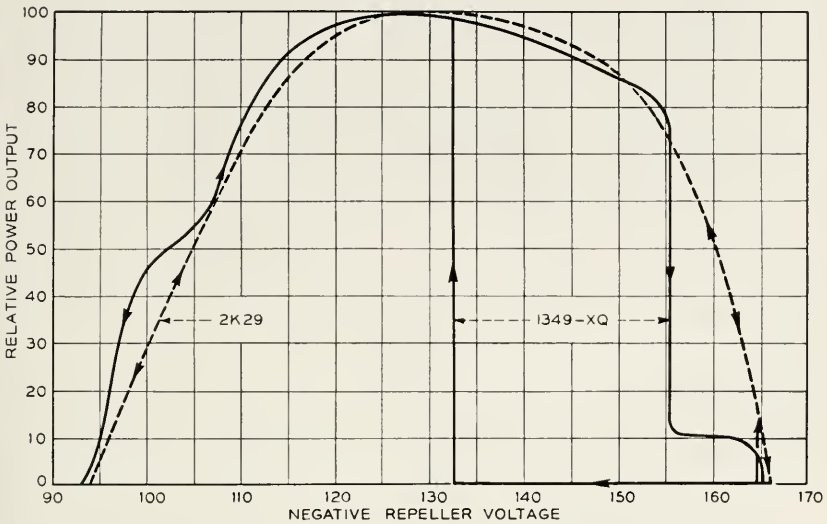


Fig. 66.—Use of the electron optical system shown in Fig. 65 eliminated the bad features of the repeller characteristics of the earlier 1349XQ in which the electrons were returned to the repeller region and gave the repeller characteristic of the final 2K29 (dashed line).

an impedance transformer and in some instances directly. As practice developed it became standard to design for optimum oscillator output characteristics with output line coupled to a 50 ohm resistive impedance. In the second case the adapter serves to couple the tube output line to the guide through a transducer.

A typical example of a reflex oscillator incorporating this construction is the 2K29. This tube is intended to cover the frequency range from 3400 to 3960 Mc/s. An external view of this tube is shown in Fig. 67. It will be observed that the center conductor of the output line extends beyond the polystyrene supporting bushing. Fig. 68 shows an adapting fitting which

permits the oscillator to be coupled to a fifty ohm coaxial cable. The center conductor of the tube output line projects into the split chuck, while contact is made to the outer conductor of the tube output line by the spring contact fingers F. The coupling unit can be mounted in a standard octal socket so that the tube can be coupled simultaneously to the power supplies and the high frequency circuit by a simple plug in operation. It is frequently desirable to insulate the outer conductor of the oscillator from

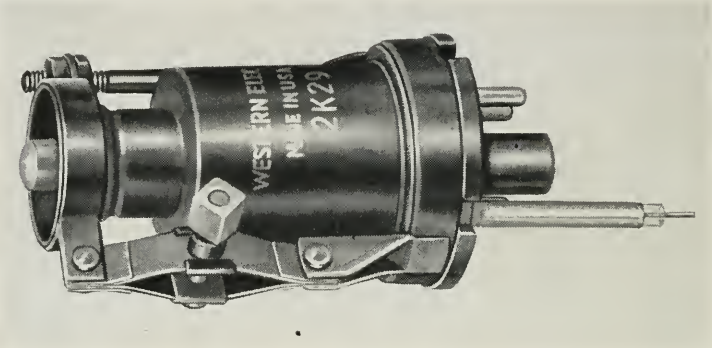


Fig. 67.—Metal reflex oscillator with enclosed resonator designed for operation from 3400 to 3960 mc/s.

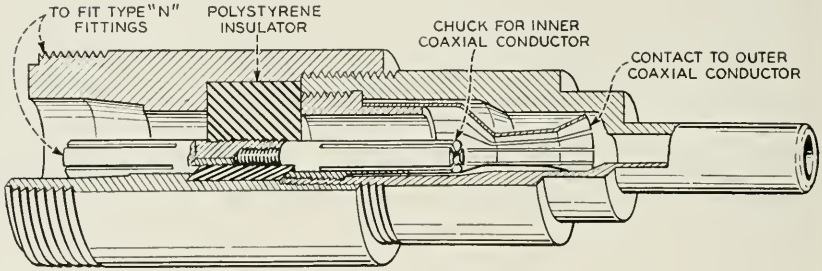


Fig. 68.—A transducer for connecting the output lead of the 2K29 to a 50 ohm cable.

the line for direct voltages while maintaining the high frequency contact. This can be accomplished with either an insulating sleeve of low capacitance or a modification of the design which incorporates a high frequency trap in the outer conductor. In some instances it is necessary to insulate the center conductor of the tube from the line for direct voltages. This can be done with a concentric condenser. The characteristic impedance of the section of the coupler may be made the same as that of the line by maintaining the proper ratio of the diameters of the condenser and the outer

conductor or may be arranged so that the condenser serves simultaneously as an impedance transformer to transform the impedance of the line to that required for best performance of the oscillator. A general discussion of the problems involved in such coupling designs will be given in a later section.

One of the primary considerations in the design of a reflex oscillator is the choice of resonator characteristics. The various controlling factors have been outlined in previous sections. In Section X it was shown that the power output and the electronic tuning optimize at different gap transit

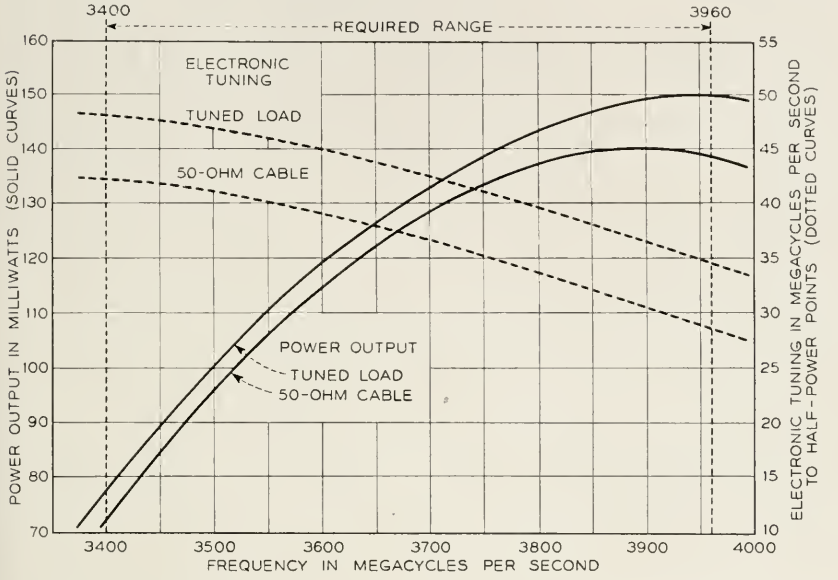


Fig. 69.—Power output and electronic tuning vs frequency for the 2K29 reflex oscillator The solid lines give the performance into a load adjusted for optimum power output at each frequency. The dashed lines show the performance obtained when the tube is coupled to a 50 ohm load by means of the fitting of Fig. 68.

angles. It is therefore necessary to compromise with the ultimate use in mind. In a beating oscillator for a radar receiver, uniformity of the electronic tuning is of greater importance than uniformity of power output, since an adjustment of the coupling to the crystal permits some variation of the latter quantity. Hence, the resonator characteristics are chosen to provide as nearly uniform electronic tuning as possible. Fig. 69 shows the electronic tuning and power output characteristics for the Western Electric 2K29 tube. These are shown for two conditions. The solid lines show the power output and electronic tuning measured into an adjustable load which

made it possible to present to the oscillator at each frequency the admittance into which the oscillator delivered maximum power output. The solid lines show the power output and electronic tuning over the band with the oscillator connected to a load presenting a resistive 50 ohm impedance to the coupling unit of Fig. 68. The problems involved in obtaining such performance will be outlined in the next section.

E. *Broad-Band Reflex Oscillators—The 2K25*

As experience with the design of radar systems and components developed and as a better understanding of the operation and limitations of the individual components was achieved, a great deal of effort was directed toward simplifying and making more reliable the number of adjustments required to optimize the performance of a system. As an illustration of this problem as related to the beating oscillator, the development problem of the 2K25 will be described. As the number of radar systems in the three centimeter range increased it became apparent that to avoid self-jamming it would be desirable to assign frequencies to various sets operating, for example, in a fleet unit. Secondarily, the over-all band of the three centimeter range was widened to cover from 8500 to 9660 Mc/s. Prior to this the 723A/B had been essentially a spot frequency oscillator and had been primarily tested as such. As was so frequently the case with tubes for military requirements, it was desired that the ultimate tube be interchangeable with an existing tube, in this case the 723A/B, and hence the improvements had to be effected within its framework.

Changes in the electronic design from that of the 723A/B produced an improved performance in the 2K25. These were a modification of the electron gun which increased the effectiveness of the electron stream and the elimination of a resonance of the region containing the electron gun which coupled with the resonant cavity and in some cases impaired the performance over the wide band. Beyond this the problem concerned the determination of an output coupling system which would provide the desired properties. This will be described in detail.

One of the most serious difficulties which occurred in early radar receivers arose from the general failure to appreciate the effect of the load impedance on the performance of an oscillator. This problem has been discussed in Section IX. In early radar receivers the method of coupling the beating oscillator to the crystal was dictated mainly by mechanical convenience rather than electrical considerations, and as a result most of the discontinuities of performance due to bad load conditions which are discussed in Section IX occurred at one time or another in most of the systems. For instance, users were much surprised to find that a beating oscillator which could be tuned to frequencies both above and below that required for

reception of the signal might yet fail to oscillate at the desired frequency. The convenience of simple replaceability of the local oscillator became in this instance a cross for its designer, since an exchange of tube would frequently eliminate the effect. This led to the obvious conclusion on the part of the user that the oscillator was defective, in spite of the fact that the presumably defective tube had passed the test specifications and would operate satisfactorily in some other receiver. The plain fact, in the light of later knowledge, was that the tubes were being improperly used, so that the usual range of manufacturing variations was not tolerable.

The appreciation of this fact led to a new approach to the problem of coupling the beating oscillator to a waveguide. In early designs the oscillator was decoupled from the load by withdrawing the probe from the waveguide. This presented the oscillator with an uncontrolled admittance with disastrous results in many cases. The new approach, proposed by the group working with Dr. H. T. Friis at the Bell Telephone Laboratories, was that of designing the receiver so that the beating oscillator could be operated into an essentially fixed impedance. The crystal was in this case loosely coupled to form a part of this load, so that variations in its impedance and the impedance looking toward the TR tube were largely masked. A great many further refinements in the design of the receiver have since been proposed, but this basic principle of defining the load into which the oscillator is required to operate is fundamental to all. In the interests of simplicity of use it appeared to be desirable to endeavor to pre-plumb the coupling of the oscillator to the wave guide. The tube designer in this instance found himself perforce, as so frequently occurs in dealing with micro-waves, a circuit designer—an instructive and illuminating experience which might happily be reversed.

The wave guide coupling was made separate from the tube, both to preserve the plug in feature of the tube and to maintain its interchangeability with the 723A/B. As a further simplification it was desired that the coupling should require no adjustments. A convenient fixed load admittance to present to this coupler is the characteristic admittance of the wave guide, since this can readily be maintained fixed over a wide band. The problem, then, is the apparently straightforward one of transforming the guide admittance to the admittance which the oscillator requires for optimum performance. Actually, the problem is complicated by the fact that the optimum admittance will vary throughout the band. The electronic admittance varies with frequency even for a fixed drift angle, because the modulation coefficient of the gap varies as the oscillator is tuned. The losses of the resonator also vary with frequency, both because of skin effect, the depth of penetration of the high frequency currents changing with frequency, and because the circulating currents in the resonator are a function

of the effective capacitance of the gap. As the capacitance increases in tuning to lower frequencies the I^2R losses therefore increase.

The objective of the coupling design may not be to obtain maximum power output at all points in the band but rather to obtain uniformity of electronic tuning and power output. An additional requirement in some cases is that a minimum sink margin as defined in section VIII should be maintained. This is equivalent to the problem arising in magnetron design of controlling the pulling figure.[13]

In designing an appropriate wave guide coupling a number of variables are at one's disposal. In the case of the W.E. 2K25 the variables available are, the length of the tube probe exposed within the guide, the offset of the probe from the axis of the guide, and the distance from the probe to the shorting piston in the guide. In addition, the characteristics of the output line of the tube are adjustable, and, finally, the coupling of the loop to the resonator can be adjusted. As one might expect, there is a large number of solutions with so many variables available. The most desirable solution is one which provides a low standing wave ratio in all parts of the coupling system. The method employed in the present case was to design a wave guide to coaxial transducer which would provide a smooth broad band transition from the wave guide characteristic admittance to the admittance of the coaxial line. In the ideal case, the characteristic admittance of the coaxial line to the loop should be maintained as uniform as possible. Structural considerations in the present case led to discontinuities which had to be appropriately balanced in the final transducer. Dr. W. E. Kock of the Bell Telephone Laboratories has given an expression which, for thin probes, relates the probe length, the offset and the distance of the backing piston when given the characteristic admittance of the coaxial line and the dimensions of the guide between which a match is required. Once such a transducer has been obtained, the admittances which must be presented to it in order to obtain maximum power from the oscillator are measured over the band. From such measurements it is then possible to determine the corrections in the loop size and in the transducer to obtain a given sink margin throughout the band. This last step actually involves a certain amount of cut and try in an effort to obtain satisfactory performance in all respects. Figure 70 illustrates the transducer developed for the W.E. 2K25 oscillator for use with 1″ by $\frac{1}{2}$″ wave guide. All tests made on this tube are specified in terms of operation in this coupler and with a load having the characteristic impedance of the 1″ x $\frac{1}{2}$″ guide presented to the coupler.

Figure 71 shows a performance diagram for a typical W.E. 2K25 oscillator operating in the coupler of Fig. 70. The reference plane for the diagram is

[13] J. B. Fisk, H. D. Hagstrum and P. L. Hartman, "The Magnetron as a Generator of Centimeter Waves", B. S. T. J. Vol. XXV No. 2, pp. 167–348 (April, 1946).

not the plane of the grids of the oscillator but is instead a more accessible reference plane external to the tube, in this instance the plane through the wave guide perpendicular to its axis which includes the tube probe. It will be observed that the sink margin in the case illustrated was equal to 5.5. At the frequency at which this diagram was obtained, the minimum sink margin permitted by the test specification is 2.5. The variation in this margin results from a variety of causes. As shown in Section III the sink margin is determined by the ratio of the total load conductance to the small signal electronic conductance. The total load conductance consists of the

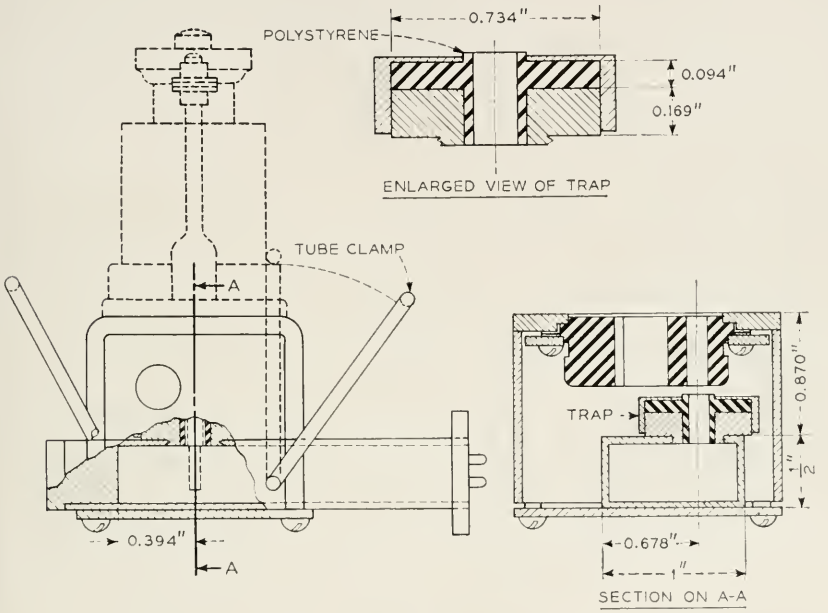


Fig. 70.—A broad band coupling designed to connect the 2K25 to a 1″ x ½″ wave guide

conductance representing the resonator losses and the conductance arising from the wave guide load transformed through the coupling system. Hence, the coupled load will be subject to variations in the loop dimensions, the characteristics of the coupling line and the transducer. The resonator loss will differ from tube to tube because of the variation in the heating of the grids and resonator by the electron stream, and there will be variations arising from other causes. The electronic conductance varies from tube to tube primarily because of the spread in beam current and secondarily as a result of such factors as variations in the modulation coefficient of the gaps, non-uniformities in the drift space causing a spread in the transit time and

the like. The sum total of these variations necessitated the maintenance of an average sink margin of 6.7 times in order to insure a minimum of 2.5.

Figure 72 illustrates further characteristics of the 2K25 oscillator and coupling. These data are the results of standing wave measurements looking towards the coupler with the tube inoperative. From such "cold test"

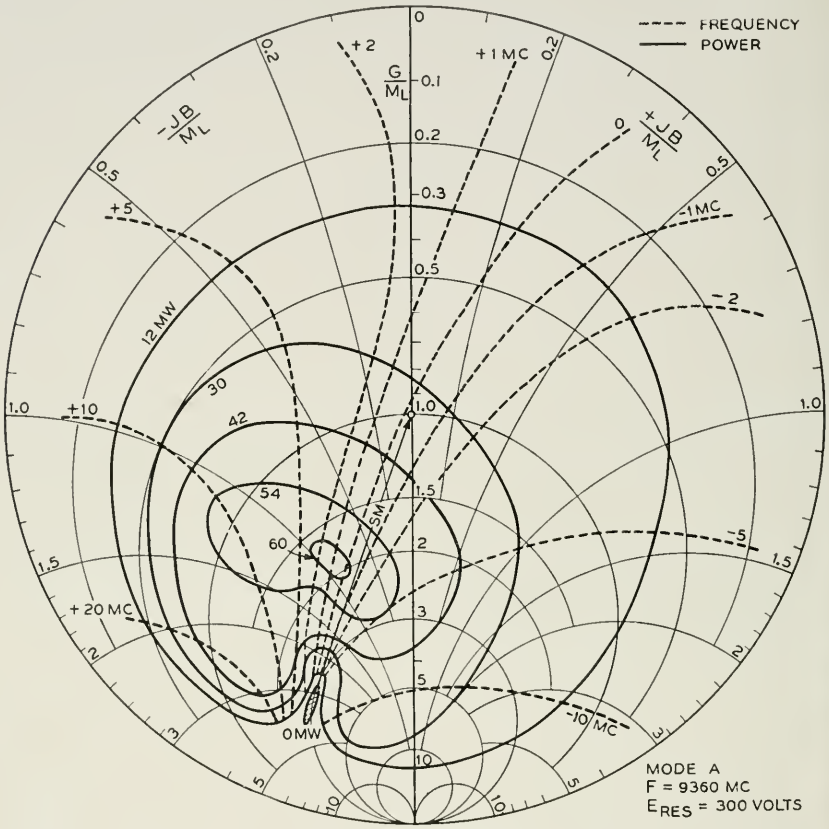


Fig. 71.—A Rieke diagram for the 2K25 connected to the load by the coupling of Fig. 70. This diagram was obtained for a repeller mode having a drift angle of 15.5 π radians at a nominal frequency of 9360 mc/s.

measurements one may determine the intrinsic Q_0 of the resonator and the external Q_E. The former is a measure of the resonator losses while the latter is a measure of the tightness of coupling of the oscillator to the load. The values of Q_0 measured from a cold test have little significance in an oscillator in which heating of the resonator and especially of the grids is appreciable. This is particularly true of oscillators tuned by variation of

the capacitance of the interaction gap. It is possible to make hot tests in which the thermal conditions of operation are established without the interaction effects of the beam, but these measurements are not available for the 2K25. The external Q_E is not affected, at least to a first order, by thermal effects in the resonator. The third curve of Fig. 72 shows the ratio of the power delivered when a matched load is coupled to the coupler to the power delivered to a load which presents optimum impedance to the oscil-

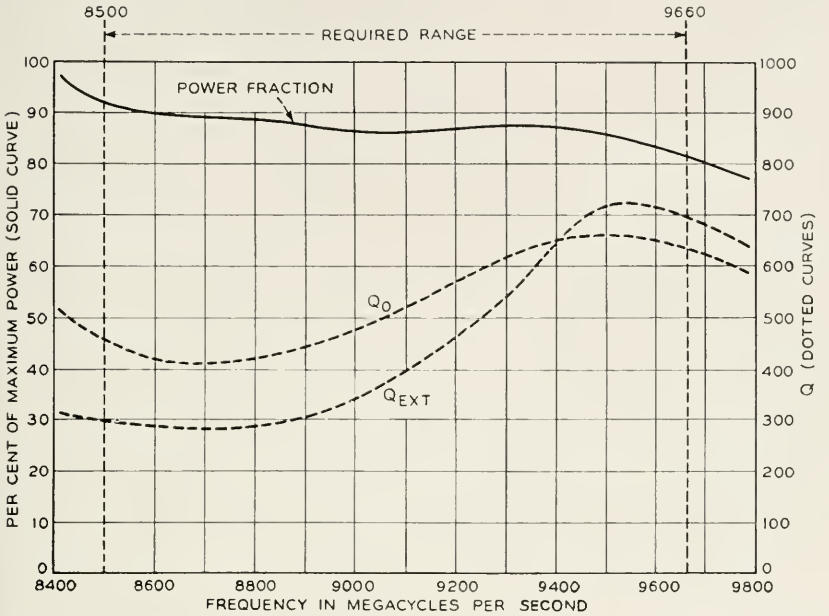


Fig. 72.—Variation of the percentage of maximum power output delivered, unloaded Q, Q_0, and external Q, Q_E, as functions of frequency for the 2K25 when coupled to the characteristic admittance of the 1″ x ½″ guide with the coupling of Fig. 70. The power variation is for a mode having 15.5 π radians drift.

lator. These data are given for the normal operating repeller mode as discussed below.

It was pointed out early in this work that the available power output and electronic tuning have a contrary variation with respect to the number of cycles of drift in the repeller space. Consequently, this is one of the most important and exasperating parameters of the tube. Fig. 73 is a diagram illustrating the characteristics of a typical W.E. 2K25 oscillator. The abscissa is the repeller-cathode voltage which, for a fixed resonator voltage, determines the drift angle. Thus, as this voltage is made increasingly negative, successive modes of oscillation appear, corresponding to consecu-

tive decreasing values of n. Our best determination for the value of n for each mode is given. The base lines are displaced vertically on a uniform wavelength scale, so that the variation of repeller voltage with wavelength is indicated. The power output increases with decreasing values of n but the half power electronic tuning for each repeller mode has a contrary variation. The repeller mode chosen as providing the best compromise between power output and half power electronic tuning is the $7\frac{3}{4}$ cycle mode. The design of the coupling unit and all the primary characteristics of the tube are based on the use of this mode. It will be observed that repeller

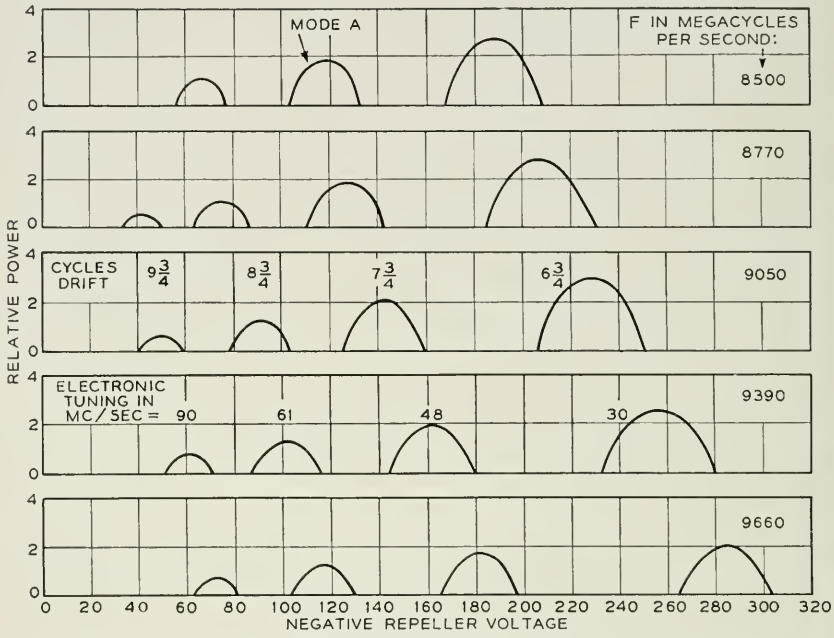


Fig. 73.—Operation of the 2K25 in various repeller modes and at various frequencies when connected to characteristic admittance of a 1″ x ½″ guide by the coupling of Fig. 70.

modes having n values less than 6 do not appear in Fig. 73. For values of n = to 0, 1, 2 and possibly 3, this is because the conductance representing the resonator losses is in excess of the electronic conductance. For the values of $n = 4$ and 5 the coupled load conductance plus the resonator loss conductance in the specified transducer are in excess of the electronic conductance. Conversely for modes having n values in excess of 7 the coupling is weaker than desired.

Fig. 74 illustrates the broad band characteristics for a typical W.E. 2K25 tube operating in the coupling of Fig. 70 into a matched load. In Fig. 74

are shown the power output, half power electronic tuning, and sink margin as functions of frequency for the $7\frac{3}{4}$ cycle mode.

F. *Thermally Tuned Reflex Oscillators—The 2K45*

The trend toward the simplification of radar systems to the fewest possible adjustments, coupled with the ever present possibility of enemy jamming, led to the attempt to produce a system which was described as a single knob system. The ultimate objective of such a system was ability to shift the frequency of the transmitter at will with a single control. This puts the chief burden on the receiver, which must automatically track with the

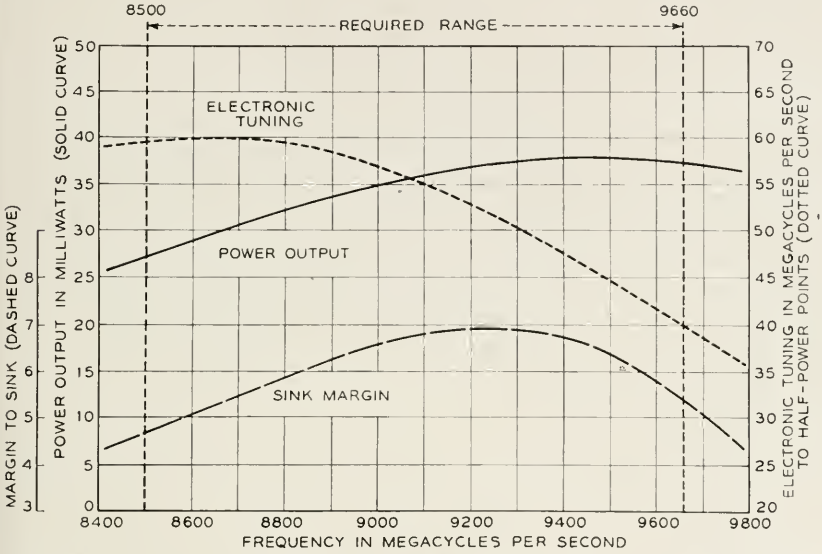


Fig. 74.—Variation of electronic tuning, power output and sink margin with frequency for the 2K25 in a repeller mode having 15.5 π radians drift. Characteristic admittance load and coupling of Fig. 70.

transmitter. The problem is much simplified by designing as many of the components as possible so that retuning is not required when the frequency is shifted within the requisite band. In the case of the beating oscillator it was necessary to devise a mechanism which would permit rapid automatic control over a frequency range of 1160 mc. This range was many times in excess of any immediately realizable electronic tuning range. It is of course apparent that such a method of tuning will also lend itself readily to use in many applications in which, although the transmitter frequency is nominally fixed, the system is required to operate under such extreme conditions that the sum total of the possible frequency deviations is in excess of the available

electronic tuning range. Examples of such systems are installations in high altitude air craft, in which wide variations in temperature and pressure may be expected.

It was highly desirable to have the frequency control electrical. One means of obtaining such a control is through motion of the resonator grids produced by the thermal expansion of an element heated electrically. A step in this direction was taken in the Sperry Gyroscope Company 2K21 oscillator in which the resonator was tuned by the thermal expansion of a strut heated by passing a considerable current through it.

At the Bell Telephone Laboratories the design for a thermally tuned beating oscillator was based on a method which permitted the control of a small current at a high voltage. In general, controlled high voltages are easily available both from power supplies and from control circuits. Further, it seemed desirable that the control of the heating should require no power. These considerations suggested that the heating of the thermal tuning element be accomplished by electron bombardment. Through the use of a negative grid to regulate the bombardment, the tuning control became a pure voltage adjustment. The bombardment method made it possible to utilize configurations in the tuner which would have been less practical if resistance heating had been employed.

An early reflex oscillator incorporating these ideas was the Western Electric 2K45 vacuum tube. Fig. 75 shows an external view of the tube which, except for the output coaxial line, looks like a forshortened 6L6 vacuum tube. The plug-in feature of the earlier mechanically tuned oscillators was maintained in this oscillator, which was designed to couple to the waveguide circuit through the same transducer developed for the 2K25.

Figures 76 and 77 are cross-sectional views of the 2K45 made at right angles to each other. The thermal tuning mechanism is contained in the upper part of the structure. It is a bimetallic combination consisting of a U shaped channel and a multi-leaf bow. The channel is formed of a material with a large coefficient of expansion and a high resistance to slow permanent deformation or creep at elevated temperatures. At the ends of this channel, tabs bent down at right angles to the channel axis, provide rigid vertical support for the channel without interfering with axial expansion. These tabs are connected to the resonator, which in turn is supported by the vacuum envelope as closely as possible in order to minimize the thermal impedance of the path. This connection also serves to cool the channel ends. The multi-leaf bow is welded to the channel at its end. The leaves are made of a material having a low coefficient of expansion and, as they are fastened to the channel at its ends, they remain cool and do not expand appreciably as the channel is heated. The purpose of the multi-leaf con-

struction of the bow is to reduce the internal stresses produced by bending during the tuning action.

A cathode, control grid and a pair of focussing wires are supported by micas in a position facing the open side of the U channel. The channel serves as an anode for the cathode current, which is controlled by the grid. The focussing wires beam the cathode current into the anode. The grid is proportioned so that under all operating conditions it remains negative, and the control system need supply no power to it.

The heating of the channel by the electron bombardment causes it to expand with a large differential with respect to the bow. As a result the bow flattens out and its center moves toward the channel. The purpose of this construction is to provide a magnification of the expansion of the

Fig. 75.—An external view of the W.E. 2K45—an early thermally tuned reflex oscillator designed by the Bell Telephone Laboratories.

channel which by itself would provide insufficient motion. The cross member welded to the center of the bow and the vertical struts transmit the motion of the bow to the diaphragm which supports one of the cavity grids. The action is illustrated in Fig. 78 which shows a series of X-ray views of an operating tube. Thus, the first view shows the conditions for no power applied to the tube, the second, for the tube operating but with the tuner grid biased to cut-off. The following views show the behavior with progressive increases in the power into the tuner channel.

The ramifications in the design of a thermal tuner are many and the possible configurations of the mechanism will depend greatly on the individual requirements of the application. It is possible, however, to lay down some basic principles. These are concerned with positiveness of action and speed of tuning. With regard to the first, it may seem anomalous to speak of

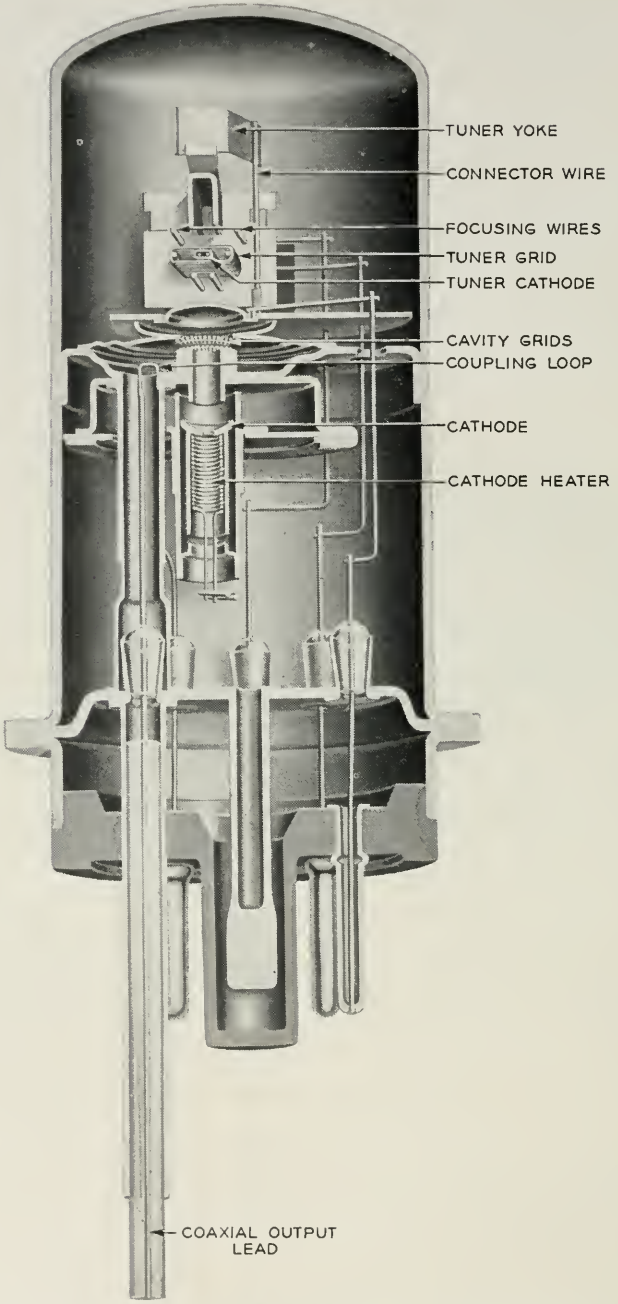


Fig. 76.—Internal features of the W.E. 2K45: section through the output lead and normal to the tuner cathode and strut.

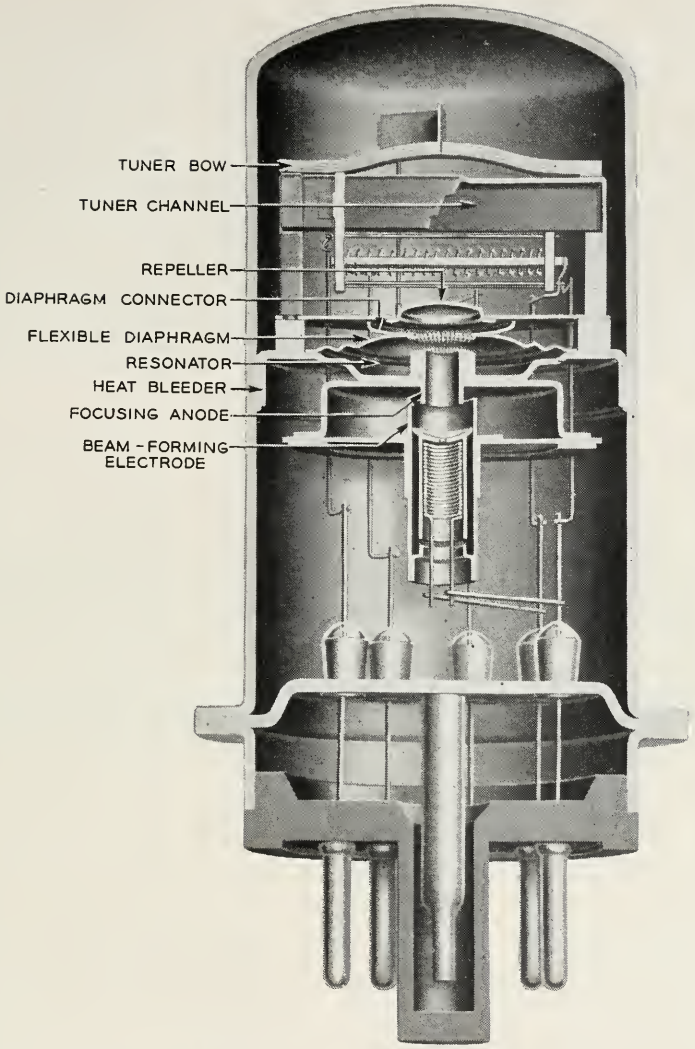


Fig. 77.—A section of the W.E. 2K45 cut parallel to the tuner cathode and strut.

positive action in a device which has thermal inertia. What is actually meant, however, may be best indicated by the following: Let us suppose that some power, P_a, must be dissipated in the tuner in order to hold the oscillator at a frequency, f_a. Now, suppose that the tube is operating with power less than P_a and that we are required to switch to frequency f_a. In order to switch rapidly, we apply power in excess of P_a to the tuner until the frequency has reached f_a, at which time we switch to the power P_a required

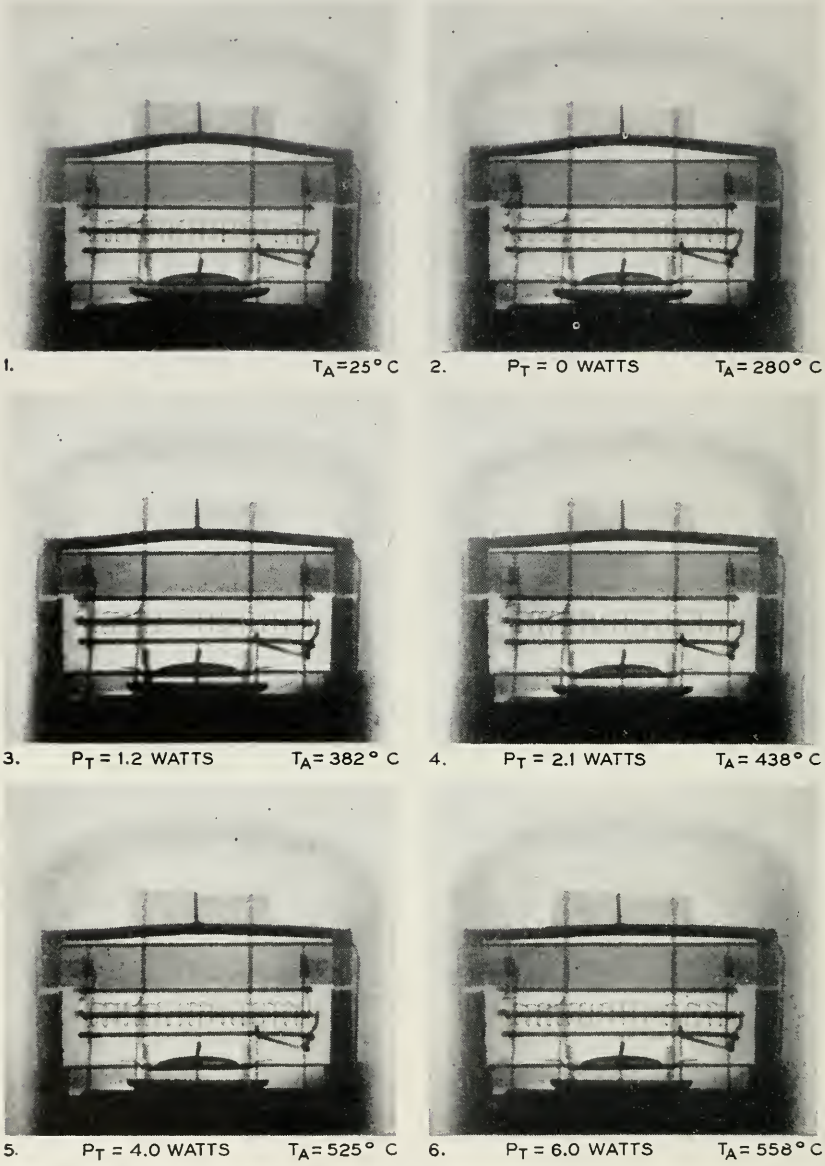


Fig. 78.—A series of x-ray photographs of the thermal tuning mechanism of an operating 2K45. These show the motion of the bows for successively increasing power inputs to the tuner channel.

to sustain f_a. We say that the control is positive if no overshoot occurs, i.e., if the frequency does not continue to change for a time and then return to

f_a. We might equally well have illustrated this by an example in which the tuner was cooling. In order to avoid overshoot it is necessary that no thermal impedance exist between the heat source and the expanding element. Thus, as an example of a tuner in which overshoot will occur, one may cite an expanding strut in the form of a tube heated by a resistance heater internal to the tube. In order to transfer heat from the resistance heater to the tubing the former must of necessity operate at the higher temperature. Hence, in the example given above, when the power is switched to the sustaining value heat will continue to be transferred to the tubing until the heater and the tubing arrive at the same temperature. To minimize the thermal impedance the heat should be generated within the body of, or on the surface of, the expanding element. The resistance heating by current passing through the expanding element illustrates the first case and heating by electron bombardment the second.

The second principle is quite obvious when once stated. If a rapid shift in frequency is to be obtained at any point within the required tuning range, then the potential tuning range must be considerably in excess of that required. Thus, if the tube is operating near one of the required frequency limits and one demands that it go to the limit, the shift will progress very slowly in the absence of excess range. If, however, it is possible to overdrive the limit, the time required will be materially shortened. On the basis of actual tube design this requires that the safe maximum or full on power into the tuner must be considerably in excess of the power required to hold the tuner at the frequency band limit nearest the full on condition. The tuning mechanism must be capable of continuous operation under the full on condition in case this accidentally persists. Further, the power required to hold the tuner at the other end of its band must be considerably in excess of zero in order that rapid cooling may occur near this limit. It is not necessary that the tuner produce motion for power inputs outside the band limits; the essential condition is that the rates of heating and cooling near these limits should have values considerably greater than zero.

It is always possible to purchase heating speed by the expenditure of power in available over-drive. The cooling speed, on the other hand, is controlled by the temperature difference between the source and sink, the heat capacity of the tuner and the mechanism of cooling. Two methods of cooling are available, conduction and radiation. For small amounts of motion and in circumstances where large forces are not required, conduction cooling can provide a satisfactory answer. In cases in which a large amount of motion is required, as in the 2K45, conduction cooling imposes a number of serious restrictions. The expanding element must be made from a material having a large coefficient of expansion and necessarily must be long. Unfortunately, alloys having large expansion coefficients are very poor con-

ductors of heat. In the case of conduction the cooling rate will depend on the ratio of the length times the heat capacity divided by the cross-section. For radiation cooling the rate depends on the heat capacity divided by the radiating area and is independent of the length except as the heat capacity depends upon this factor. Radiation cooling has the advantage that it permits more freedom of structural shape and the material may be chosen on the basis of strength. In the operating range of the 2K45 the heat radiated is approximately four times that conducted away.

The automatic frequency control circuits which have been used with thermally tuned beating oscillators in radar receivers have been of a full on or off type so that they do not continuously hold the frequency at a fixed reference difference from the transmitter frequency. The reason for this in part is that a pulse transmitter gives the required reference information only during the pulse. Thus with an on-off control circuit if at a given pulse a correction of frequency is demanded the power into the tuner is turned full on or off, depending on the direction of the correction, and left in this condition until the occurrence of the next pulse which again determines the direction of the control. The result is that the frequency of the beating oscillator continually hunts about the transmitter frequency. The control system must be designed to hold the hunting deviation within tolerable limits. It is of course possible to work within narrower limits than full on or off tuner power. One advantage of full on or off control is that it results in the minimum tuning time between required frequencies since the tuning rate is at all times held to the maximum possible value.

Under certain conditions the performance of a thermal tuner may be completely described by a time constant. In general, however, this is not the case and the information required in designing a control system is concerned first with initiation of operation and second with the factors determining the magnitude of the hunting deviation. For initiation of operation, one is concerned with the time required for the oscillator to reach the operating frequency. The quantities known as the cycling times give upper limits for this quantity. The cycling times are to a certain extent arbitrarily defined as indicated in the following. There are two band limit frequencies, f_c, the limit requiring the lesser power input, P_c, and corresponding to a tuner temperature, T_c, and f_h, the limit for which the power input will be P_h, and the temperature, T_h. The cycling time for heating, τ_h, is defined and measured in the following manner. The power input to the tuner is set to P_c and held at this value until equilibrium is established. The power input is then switched to the maximum allowed value, P_m and the time interval, τ_h, between the instant of switching and the instant at which the frequency of operation has reached f_h is measured. Correspondingly, the cycling time τ_c for cooling is measured by setting to P_h until equilibrium is

established and then switching the tuner power off and measuring the interval, τ_c, until the operating frequency reaches f_c. These quantities are of importance in determining the "Out of Operation" time in case the frequency reference of the control system is momentarily lost, so that the control starts cycling in order to re-establish the reference.

While the cycling times can be taken to give an indication of the average speed of tuning, more detailed information is required to determine the hunting deviation. This demands a knowledge of the instantaneous tuning rates which will result at any point in the band when the power is switched full on or off. These rates vary through the band since the overdrive available, for example, on heating will decrease as the operating frequency approaches the limit nearest to the maximum drive.

In the following, an outline will be given of the factors which must be considered in designing a thermally tuned reflex oscillator. The 2K45 will be used as an illustration. Our first consideration concerns the time required for the tuner to heat and cool between given temperatures. In Appendix XI expressions are derived for the cycling times τ_h and τ_c. The expressions applicable to the 2K45 are:

$$\tau_h = \frac{CT_m}{2KT_m^4} [F_1(T_{rh}) - F_1(T_{rc})] \tag{13.1}$$

$$\tau_c = \frac{CT_0}{2KT_0^4} [F_2(T_{sc}) - F_2(T_{sh})] \tag{13.2}$$

where the symbols are defined in the appendix. The functions F_1 and F_2 are plotted in terms of the reduced temperatures, T_r and T_s in Figs. 79 and 80. In the analysis conduction cooling is neglected and it is assumed that the whole of the expanding element operates at the same temperature. Because of these limitations the theory is largely qualitative. It will be observed that the cycling time, τ_h, is proportional to the ratio of the heat energy stored in the tuner at the maximum equilibrium temperature to the rate of loss of energy at this temperature. It is therefore apparent that this equilibrium temperature should have the maximum possible value, and also that the heat capacity of the tuner should be kept to a minimum. Assuming for simplicity that the frequency of oscillation is proportional to the temperature, so that a given temperature difference is proportional to the frequency, one sees by examining the function F_1 that it is desirable to keep the reduced temperatures T_{rh} and T_{rc} small compared to 1. Under these circumstances, the cycling time τ_h will have its minimum value and will be more or less independent of the reduced temperatures. If we examine the expression for the cycling time for cooling, τ_c, we observe that this is proportional to the ratio of the heat stored in the tuner at the equilibrium

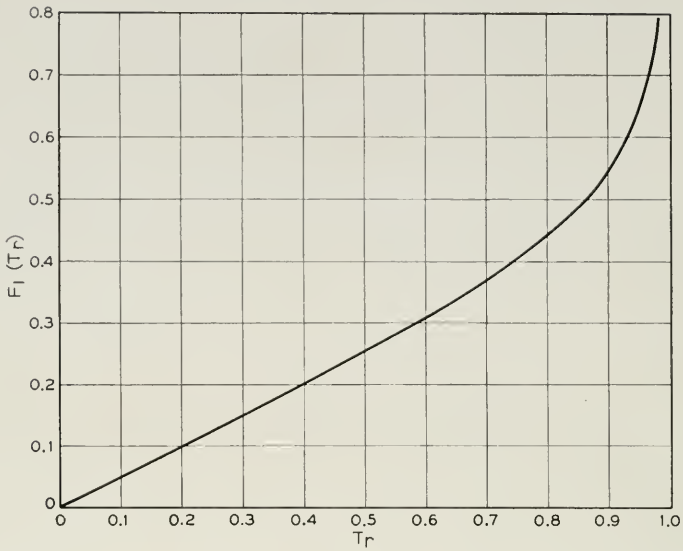


Fig. 79.—A plot of a function used in determining the time required for the temperature of a thermal tuner to rise between two given values when the tuner is cooled by radiation alone. The abscissae are given in terms of a reduced temperature.

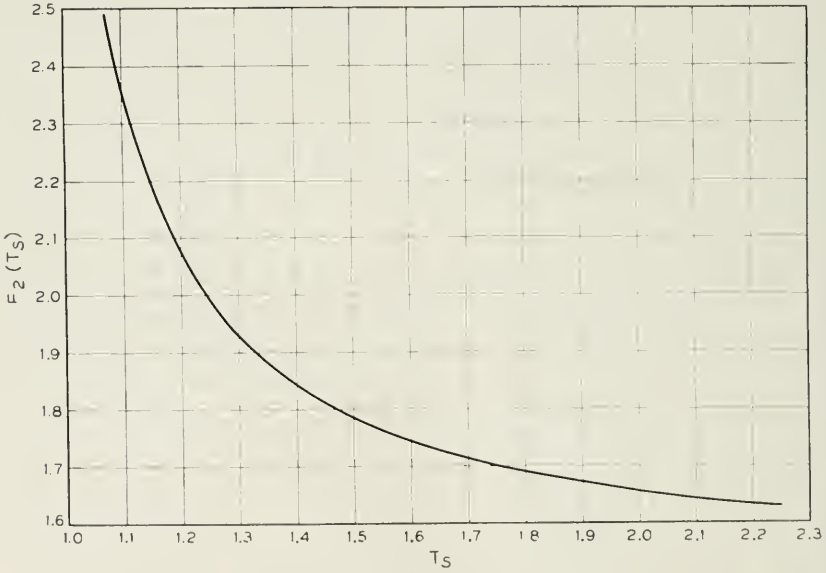


Fig. 80.—A plot of a function used in determining the time required for the temperature of a thermal tuner to fall between two given values when cooled by radiation alone. The abscissae are given in terms of a reduced temperature.

temperature T_0 corresponding to zero bombardment power divided by the rate of heat loss at this temperature. This indicates a rather paradoxical result, that the temperature for zero bombardment power should be as high as possible. This arises from the dependence on radiation cooling. We are limited in setting the value of T_0 by the form of the function F_2, which requires that the reduced temperatures T_{sc} and T_{sh} should be very large compared to 1. Since the true temperatures corresponding to these reduced values must simultaneously be small compared to T_m, it is apparent that we are not completely free to make T_0 large, and a compromise must be worked out.

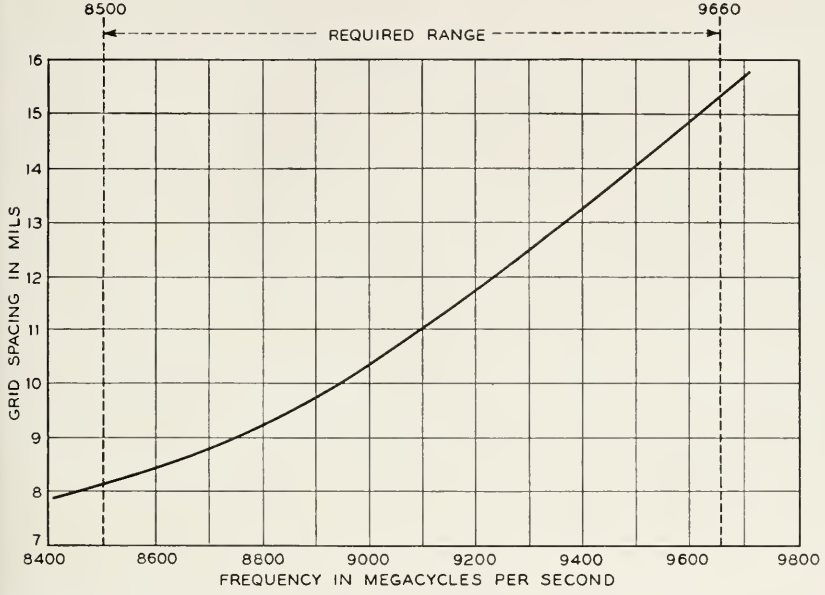


Fig. 81.—Variation of the frequency of resonance vs gap displacement for the W.E. 2K45 resonator. The vertical lines show the required tuning range.

In determining a practical design for a thermal tuner, the first characteristic which must be known is the variation of the resonant frequency of the oscillator cavity as a function of gap displacement. It is apparent that, for the highest speed of tuning, the rate of change of frequency with gap displacement should have the maximum possible value. However, this tuning characteristic is dictated by the performance requirements of the tube as an oscillator and hence is not available as a variable in the design. Fig. 81 shows the variation of frequency with gap displacement for the 2K45 resonator. The required range is indicated.

When the required motion is known a choice may be made of a mechanism

for a tuner. There is a certain amount of arbitrariness in choosing the limiting dimensions of the tuner. If the expanding element is to be short it is necessary either to operate over a very wide range of temperatures or else use some mechanical means of amplifying the motion obtained over a more limited range. Since the more limited is the required temperature range the greater is the tuning speed, it is obviously advantageous to use mechanical amplification. As previously pointed out the electron bombardment method of heating and radiation cooling are especially suitable to such a mechanism because of the freedom of design they permit. Previous dis-

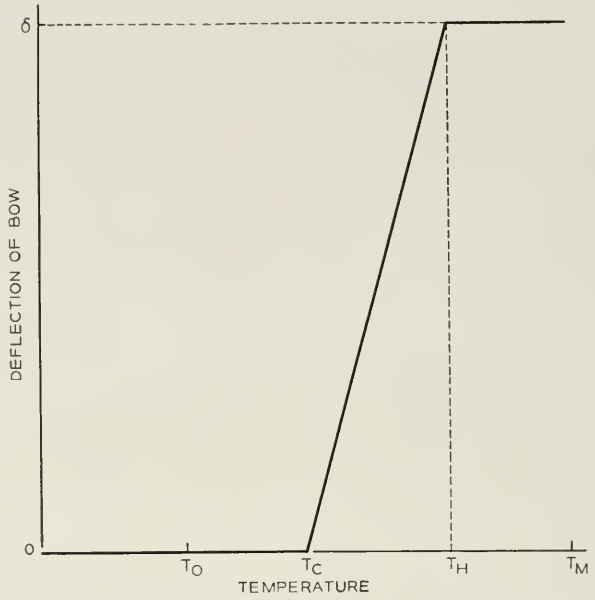


Fig. 82.—The ideal type of deflection vs temperature characteristic desired for a thermally tuned oscillator. The motion δ is that required to shift the resonant frequency of the cavity through its required band.

cussion has shown that the temperatures corresponding to zero and maximum power input must be separated by wide margins from the temperatures corresponding to the limits of the tuning range. Any motion which occurs in these margins is unnecessary and in general undesirable. Ideally, the tuning mechanism should have a characteristic as shown in Fig. 82. The type of tuning mechanism chosen for the 2K45 is a first approximation to such a characteristic, as is shown in Fig. 83, which gives a family of characteristics corresponding to various initial offsets of the bow for a given length of bow. The bows are formed to a sinusoidal shape. This structure gives

a large mechanical amplification of the expansion of the tuner. The tuner is made as long as will fit conveniently into the tube envelope. Further arbitrary decisions are required with regard to the power which can be ex-

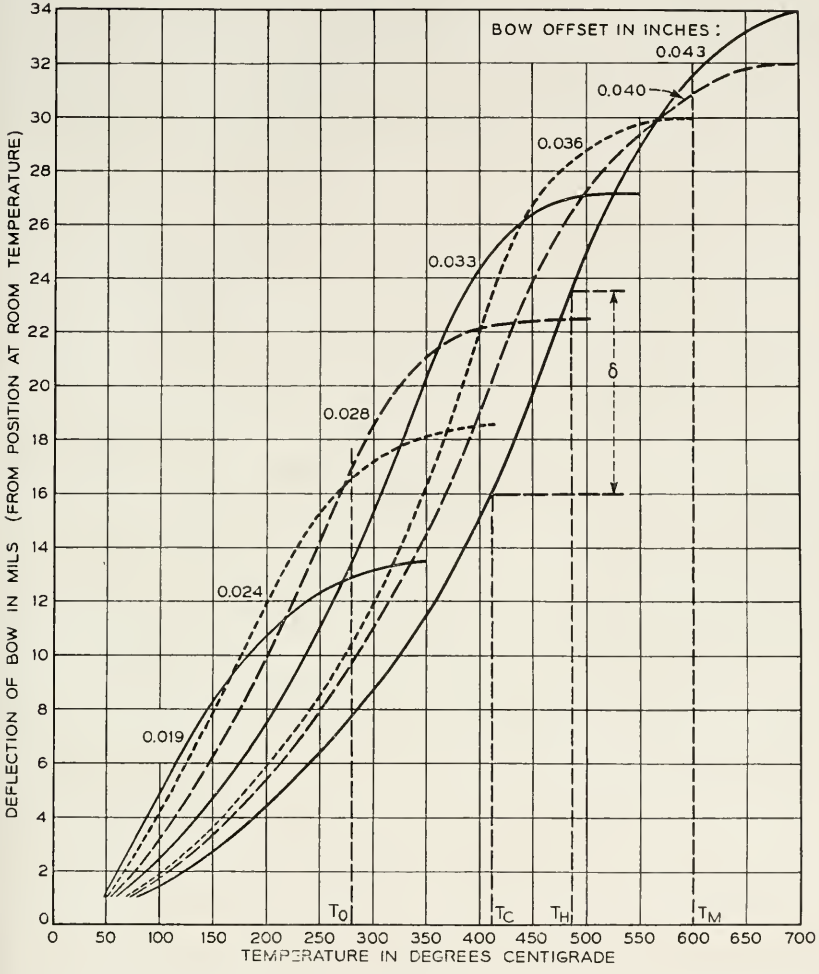


Fig. 83.—A family of deflection vs temperature characteristics for the type of tuning mechanism used in the W.E. 2K45. The parameter is the initial offset of the bows from the channel at room temperature.

pended in operating the tuner. With such decisions made, the tuner design can then be completed as a compromise of a great many variables. Limitations on the strength of the tuner materials at elevated temperatures deter-

mine a maximum safe operating temperature. The anode material is chosen to have a large coefficient of expansion and resistance to "creep" at elevated temperatures. In choosing the form of the tuner it is necessary to keep constantly in mind the necessity of maintaining the minimum ratio of heat capacity to radiating surface. A minimum operating temperature of the tuner is determined by heat flow to it from extraneous sources. Figure 84 shows the temperature as a function of bombardment power input for the 2K45 tuner. It will be observed that the heat from sources other than bombardment produces a considerable rise in temperature. One principal source of uncontrolled heating is radiation from the tuner cathode. This

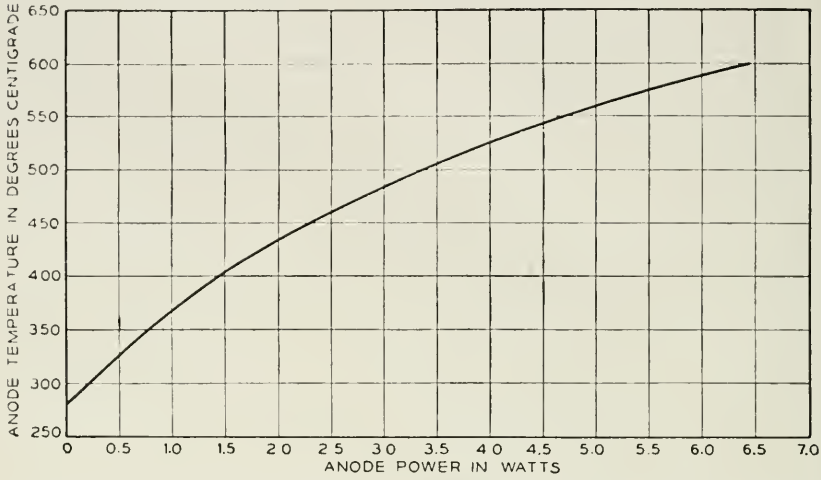


Fig. 84.—The tuner anode temperature as a function of the bombardment power. The temperature rise at zero bombardment power is caused by radiation from the thermal cathode and extraneous sources.

is minimized by keeping the cathode as remote from the anode as is consistent with the required electronic characteristics.

When the maximum and minimum operating temperatures and the length of the channel are determined the remaining problem is to determine and offset for the tuning bow which will provide the optimum tuning characteristics. We wish to obtain characteristics such that the heating time τ_h and the cooling time τ_c are approximately equal and of a minimum value. The choice of the bow offset also involves a choice of an initial gap spacing for the resonator. On Fig. 83 the boundary values for the limiting temperatures T_0 and T_m are indicated by vertical lines. With any given bow offset which corresponds to a particular tuning characteristic a limit is set on the initial gap spacing by the requirement that the total motion of the bow between

room temperature and T_m shall not exceed the initial gap spacing. From our previous analysis we know that to provide maximum tuning speed we desire to make the temperature intervals $T_c - T_0$ and $T_m - T_h$ as large as possible. For any given tuning characteristic these intervals may be adjusted by a variation of the initial gap spacing subject to the limitation just

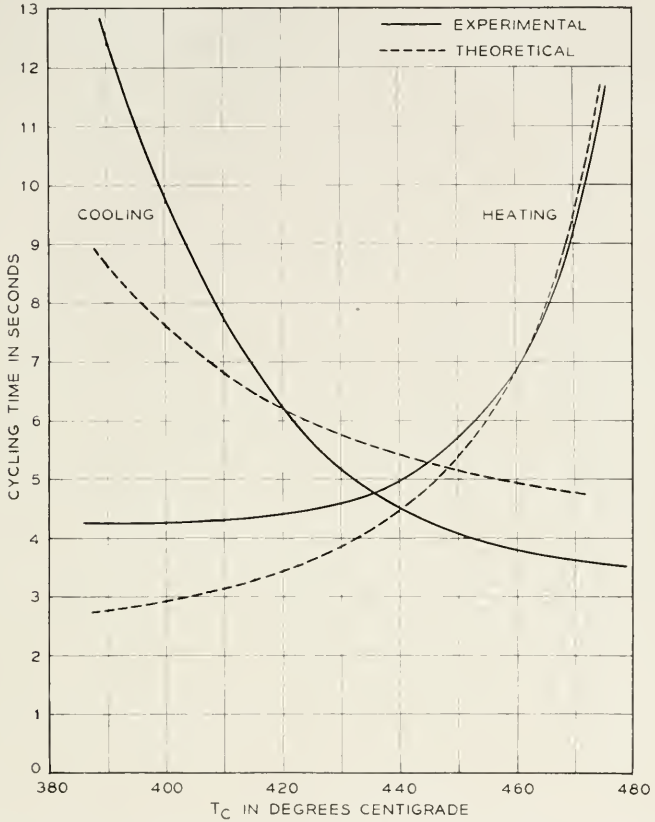


Fig. 85.—The cycling time of the W.E. 2K45 as a function of the temperature of the channel at the band frequency limit requiring the smaller tuner power. The solid lines are the experimental curves, the dashed lines are theoretical results. One point on the heating time is fitted in order to determine the heat capacity of the tuner.

imposed. Since T_c and T_h are interrelated for a given bow characteristic by the fact that they correspond to a specific increment of motion δ determined by the cavity design, we may study the effect of shifting the interval δ along the tuner characteristic by varying the initial gap offset. We may specify this in terms of value of T_c. The result of such a study is shown in Fig. 85. The optimum offset corresponds to the temperature at which the

curves for τ_h and τ_c cross over. The optimum bow offset is then the value which provides the minimum value at the crossover. From the theory given earlier it is possible to compute these curves. If we had analytical expressions for the motion of the tuner with temperature and for the variation of frequency with gap spacing it should be possible to obtain completely theoretical curves.

As a test of the theory of heating and cooling it is sufficient to use the experimental curves for the motion of the tuner with temperature and the variation of frequency with gap spacing in conjunction with the heating and cooling curves of Figs. 79 and 80 calculated from equations (13.1) and (13.2). The value of K which must be determined in order to obtain numerical values from the curves of Figs. 79 and 80 may be determined from Fig. 84 by using the relationship

$$K = \frac{P_a - P_b}{T_a^4 - T_b^4}$$

where P_a and P_b are the bombardment power inputs corresponding to anode temperatures T_a and T_b. There is no ready means for directly determining the heat capacity C. However, if one point on either the heating or cooling curves is fitted to the experimental data the value of C may be determined and the remainder of the points computed. The results of such a computation are shown by the dashed lines in Fig. 85. In view of the restrictive assumptions of a uniform anode temperature and the neglect of all conduction cooling the agreement in general form is reasonably good.

The W.E. 2K45 includes a number of advances in reflex oscillator technique over the 2K25. It will be observed in Figs. 76 and 77 that the electron optical system employed in the gun differs from that used in the 2K25. In the 2K25 a gun producing a rectilinear beam was employed. In the 2K45 the gun consists of a concave cathode surrounded by a cylindrical electrode and a focussing anode. The design of this type of gun was originated by Messrs. A. L. Samuel and A. E. Anderson at these laboratories. The design is such as to produce a radial focus beam which converges into the cylindrical section of the focussing anode. After the beam enters the focussing anode its convergence is decreased by its own space charge, and the beam passes through the grids at approximately the condition of minimum diameter. Between the second grid and the repeller the beam continues to diverge radially on the outbound and return trips. The intention of the design is that the beam shall have diverged sufficiently so that the maximum possible fraction will recross the gap within a ring having an inner diameter equal to the first grid and outer diameter equal to the second grid. Under these conditions only a small fraction of the beam will return into the cathode region, the remainder being captured on the support of the first grid after the

second transit of the gap. This tends to eliminate electronic tuning hysteresis and the repeller characteristic of the 2K45 is essentially free of this phenomenon. This gun design has the further advantage that it avoids the necessity for the first grid used in the 2K25. This eliminates the current interception on this grid with a resulting increase in the effective current crossing the gap. This type of gun also permits the design of a more efficient resonator by reducing the grid losses.

A second variation in design from the 2K25 is that in the 2K45 the second grid moves with reference to the repeller. This has the advantage of reducing the variation of the repeller voltage for optimum power with resonator tuning. The drift angle in a uniform repeller field is given by

$$\theta = \frac{k\sqrt{V_0}}{V_0 + V_R}\,\ell f \tag{13.3}$$

where ℓ is the spacing between the repeller and the second grid of the resonator. If the same drift angle θ is maintained at all frequencies in the band, then the repeller voltage must vary. If ℓ is fixed as f increases V_R must also increase in order to maintain a fixed fraction. If ℓ varies and increases as f decreases then a smaller variation in V_R is required and in the particular case that ℓ varies inversely with f the repeller variation may be made zero. Usually other requirements determine the variation of ℓ and it is not always possible to make the variation zero. In the case of the 2K45 the variation over the band is approximately half the amount which would occur if ℓ were fixed.

The output coupling and line of the 2K45 were designed so that the oscillator would provide the desired characteristics in the same waveguide adapter as designed for the 2K25. The power output as a function of frequency for a typical tube is shown in Fig. 86a. Curves A, B and C of Fig. 86 show the variation of power output with cavity tuning when the repeller voltage is set for an optimum at the indicated frequency and held fixed as the cavity tuning is varied. The frequency shift between half power points in this case is very much wider than with repeller tuning. This is a consequence of the fact that whereas in repeller tuning both the frequency and the drift time change in a direction to shift the transit angle away from the value for maximum power, with cavity tuning only the frequency changes. Moreover, the fact that the repeller to second grid spacing in this design varies with frequency tends to reduce the variation of the drift angle with frequency. The envelope of the curves A, B and C gives the power output as a function of frequency when the repeller voltage is adjusted to an optimum at each frequency.

Fig. 86b gives the half power electronic tuning as a function of frequency measured statically and also dynamically with a 60 cycle repeller sweep. The difference arises from thermal effects.

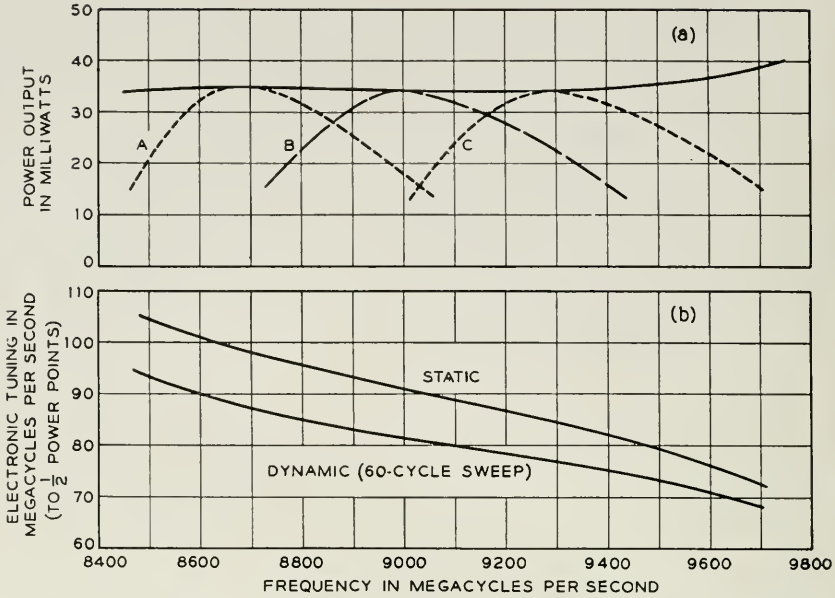


Fig. 86.—Characteristics of the W.E. 2K45 reflex oscillator. Fig. 86-a shows the variation of the power output as a function of frequency in two cases. The curves *A*, *B* and *C* illustrate the power variation with frequency when the repeller voltage is set for the optimum at the indicated frequency and held fixed as the cavity tuning is changed. The envelope of these curves shows the power variation with frequency when the repeller voltage is maintained at its optimum value for each frequency.

In Fig. 86b the electronic tuning is shown, in one case where the repeller voltage is shifted between half power points so slowly that thermal equilibrium exists at all times and in the dynamic case in which the repeller voltage is shifted at a 60 cycle rate.

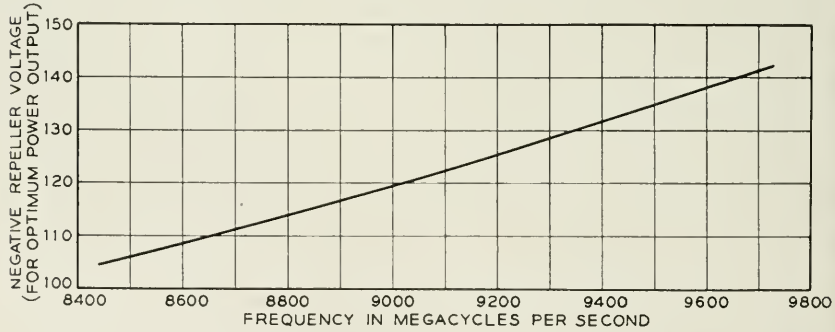


Fig. 87.—Repeller voltage for optimum power as a function of frequency for the W.E. 2K45 oscillator.

Fig. 87 shows the variation of repeller voltage for optimum power with frequency. This variation is so nearly linear that it has been proposed that

a potentiometer properly ganged with the transmitter in radar systems would provide optimum output throughout the band. This is an advantage over electronic tuning, since the signal to noise performance of the receiver depends in part on the beating oscillator power into the crystal.

In a thermally tuned tube it is necessary to provide safeguards against excessive power input to the tuning strut since this might produce a permanent deformation and impair tuner operation. A simple method for obtaining such protection is to use low-frequency cathode feedback produced by a cathode resistor. With a cathode biasing resistor of 725 ohms the grid may be held 15 volts positive with respect to the more positive end of the cathode biasing resistor indefinitely without damage to the tuner when the normal plate voltage of 300 volts is applied.

The grid control characteristics for a typical 2K45 shown in Fig. 88 were obtained while using the cathode biasing resistor. These characteristics may be given in two ways. In one case the repeller voltage is held fixed and the characteristics are given over a range between half power points. It will be observed in this case that the characteristics are discontinuous because of the electronic tuning resulting from the repeller voltage shifts between ranges. For the other case the repeller voltage is maintained at its optimum value at each frequency.

In either case, one striking feature is the essential linearity of the variation of frequency with grid voltage. This is of considerable importance in many frequency stabilizing systems and represents an advantage of thermal tuning over electronic tuning. In the case of electronic tuning, as shown in Section VII, the rate of change of frequency with repeller voltage varies rapidly as the repeller voltage shifts away from the optimum value. Since frequency stabilization is essentially a feedback amplifier problem in which the rate of change of the frequency with the control voltage enters as one of the factors determining the feedback, it is apparent that the frequency stabilization will vary as the repeller voltage is shifted. In contrast, for the case of thermal tuning, because of the linearity of frequency with grid voltage, the stabilization will be independent of the frequency. It should not be for-

2K45 Operating Conditions

	Normal	Maximum
Heater Voltage	6.3	6.8 Volts
Resonator Voltage	300	350 Volts
Klystron Current	22	30 mA
Repeller Voltage Range	−60 to −175	−350 Volts
Tuner Current	0 to 25	mA
Tuner Power		7.0 Watts
τ_h (9660-8500 *Mc/s*)	6.0	9.0 Sec
τ_c (8500-9660 *Mc/s*)	6.0	9.0 Sec

gotten, however, that thermal tuning is inherently slower in action than electronic tuning, since the latter is capable of frequency correction rates limited, for practical purposes, only by the control circuits, whereas in

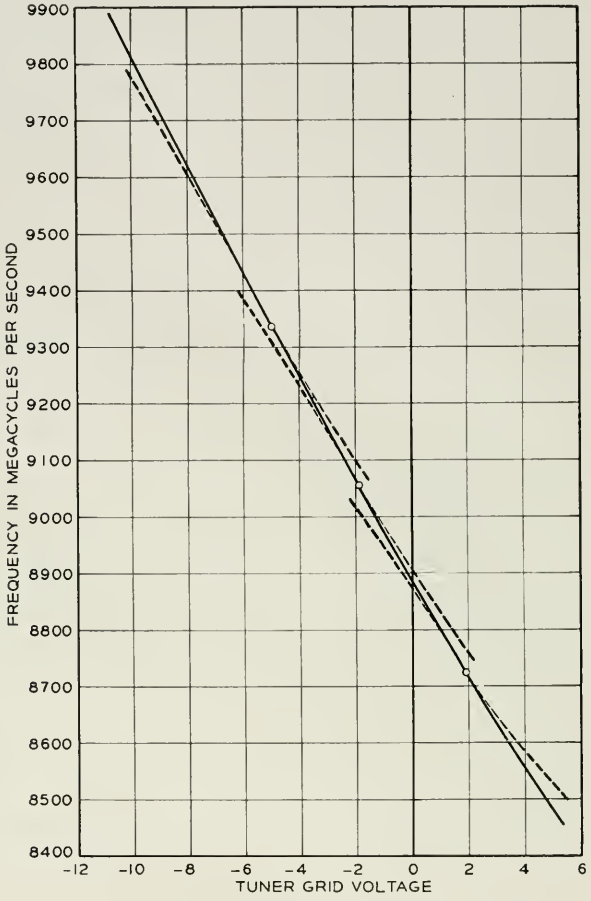


Fig. 88.—Frequency as a function of the grid voltage for the W.E. 2K45 oscillator. The dashed lines show the variation when the repeller voltage is held fixed for the mid-range value and the solid lines shows the variation with the repeller voltage maintained at its optimum value for each frequency. These characteristics apply when a 725 ohm biasing resistor is connected in series with the tuner cathode.

thermal tuning the thermal inertia of the tuning strut limits the tuning speed to rates of the order of 100 mc/sec^2 for the 2K45.

The operating conditions for the 2K45 are given in the table at the bottom of page 595.

G. *An Oscillator With Waveguide Output—The 2K50*

Late in the war it became apparent that there was an urgent need for radar systems which would permit a very high degree of resolution. Such resolution requires the use of the shortest wavelengths possible, and as a practical step development work was undertaken in the neighborhood of 1 cm.

Work at these Laboratories led to a tube which produced over 20 milliwatts and was thermally tunable over the desired frequency range by means roughly similar to those employed in the 2K45. This tube had a wave guide output. It had no grids; a sharply focused beam passed through a .015″ aperture in the resonator. The tube operated at a cavity voltage of 750 volts.

Work by Dr. H. V. Neher at the M.I.T. Radiation Laboratory resulted in a design for an oscillator using grids which operated at a resonator voltage of 300 volts. At the request of the Radiation Laboratory, the Bell Telephone Laboratories undertook such development and modification as was necessary to make the design conform to standard manufacturing techniques. This work was carried out with the close cooperation of Dr. Neher.

Figure 89 shows an external view of the tube and Fig. 90 a cross sectional view of the final structure. There are two striking departures in this tube from the designs previously described. One of these is that the axis of symmetry is no longer parallel to the axis of the envelope but instead is perpendicular to it. This construction makes possible in part the other striking feature of the tube, which is the wave guide output. A number of factors combine to make this type of output desirable and practical. The resonant cavity for a wavelength near 1.25 cm. becomes extremely small. Were loop coupling used this would necessitate a very small coupling loop and also a very small diameter output line. The small dimensions with loop coupling would require tolerances extremely difficult to maintain with conventional vacuum tube techniques. On the other hand, the wave guide used at 1.25 cm. is of dimensions (.170″ x .420″) such that a wave guide output with a choke coupling can readily be incorporated in a standard vacuum tube envelope.

The wave guide coupling is accomplished by means of a tapered wave guide which couples to the cavity through a non-resonant iris. The guide tapers in the narrow dimension only, from the iris to a circular output window. The tapered guide couples to the window by means of a circular half wave choke. The VSWR introduced by the window is 1.1 or less. External to the tube, there is an insulating fitting which permits the tube to be coupled directly to the guide by means of a second choke coupling. This makes it possible to operate the shell of the tube at a different potential

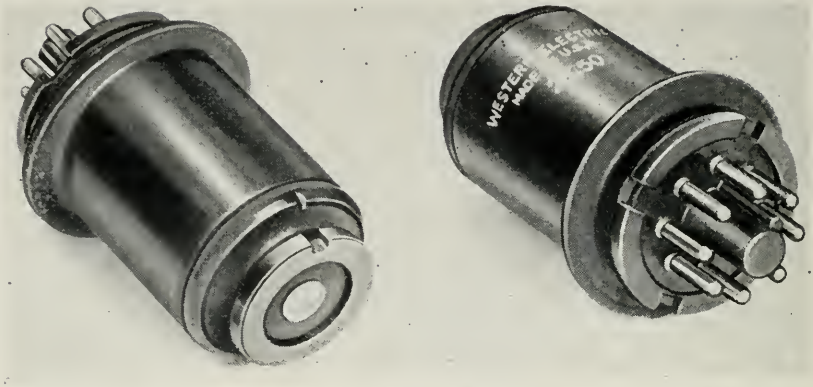


Fig. 89.—The 2K50—a reflex oscillator with thermal tuning and a wave guide output or operation in the 1 centimeter range.

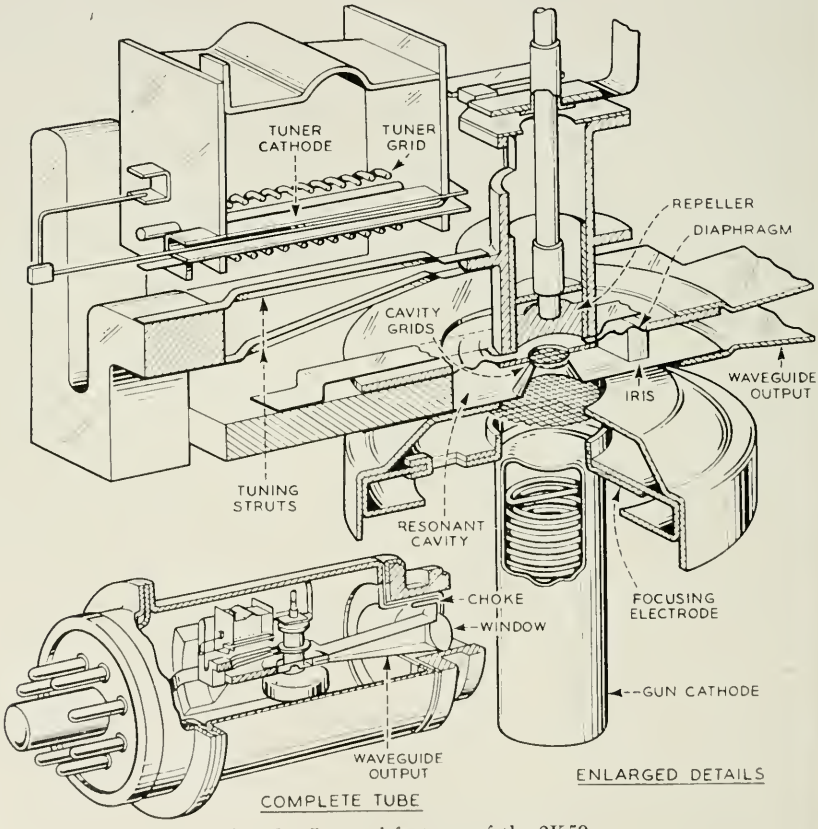


Fig. 90.—Internal features of the 2K50.

than the guide. This is desirable in a radar receiver for circuit reasons, which require that the cathode of the oscillator be at ground potential.

The iris size is a compromise chosen to provide sufficient sink margin throughout the band. An iris coupling inherently varies with frequency and provides a weaker coupling at lower frequencies. Hence, since a suffi-

Fig. 91.—A performance diagram for the 2K50 at the high frequency band limit. This diagram shows loci of constant power as a function of the admittances presented at the plane of the tube window. Admittances are normalized in terms of the characteristic admittance of the wave guide.

cient sink margin must be provided at the wavelength where the coupling is a maximum, this means that an excess of sink margin exists at the low frequency end of the band. This is illustrated by the impedance performance diagrams of Figs. 91 to 93.

The 2K50 presented a difficult mechanical problem which will be appreciated when the minute dimensions of the resonant cavity are observed in

Fig. 90. The electron optical system consists of a concave cathode, a cylindrical beam electrode, and a grid concave towards the cathode. This produces an electron beam which converges into a conical nose and through the cavity grids to the repeller space. The repeller, which returns the beam across the gap, is rigidly held in a mica supported in a cylindrical housing connected to a diaphragm which serves as one wall of the resonator. The

Fig. 92.—Performance diagram for the 2K50 at the mid-band frequency.

cylindrical housing is connected to a thermal tuning mechanism consisting of a simple framed structure in the shape of a right triangle. The base of the triangle is a heavy piece of metal brazed to the cavity block and a bleeder shoe which in turn is brazed to the bulb. One leg of the triangle can be heated by electron bombardment controlled, as in the 2K45, by a negative grid. The other leg is heated only by conduction. Since both legs are made from the same material, general heating of the structure will produce

only a second order effect. Because of the small motion required to tune the 2K50 through its range, the tuner dimensions permit reliance on conduction cooling.

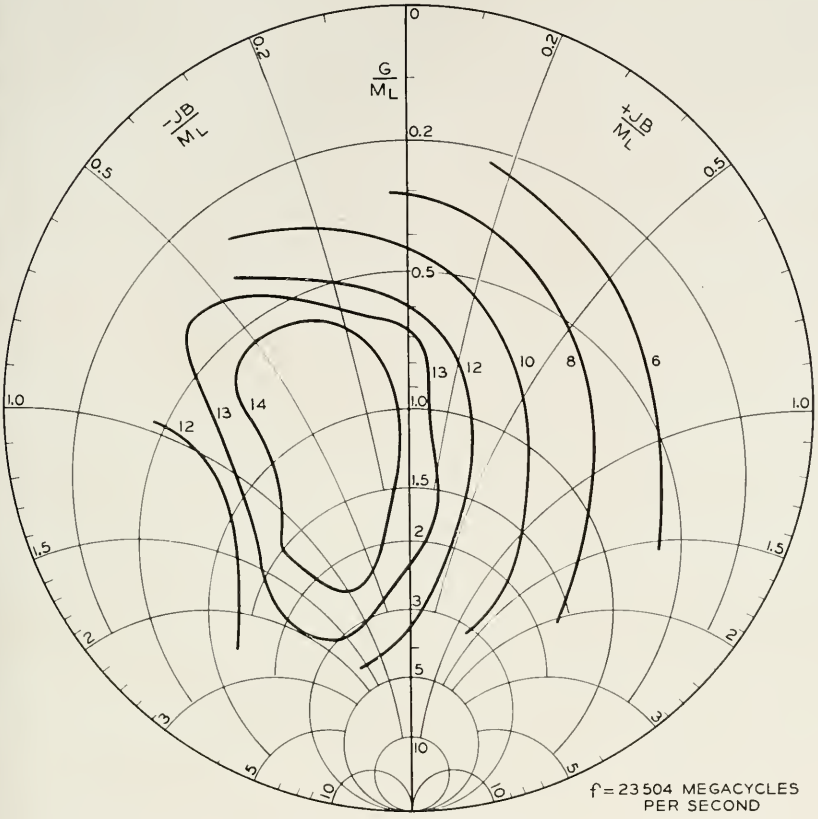


Fig. 93.—Performance diagram for the 2K50 at the low frequency band limit.

The characteristics of the thermal tuner of the 2K50 differ considerably from the 2K45. In Appendix XI expressions are given for the heating and cooling times as

$$\tau_h = \frac{c}{k} \log_e \frac{P_m - kT_c}{P_m - kT_h} \tag{13.4}$$

$$\tau_c = \frac{c}{k} \log_e \frac{T_h}{T_c}. \tag{13.5}$$

It is desirable that the cycling times should be equal. Equating (13.4) and (13.5) one obtains

$$P_m = k\,(T_h + T_c). \tag{13.6}$$

This states that the maximum permitted temperature, $T_m = P_m/k$, shall exceed the temperature T_h by the same temperature difference as that by which T_c exceeds the sink temperature. From (13.4) and (13.5) it can be seen that to minimize the heating and cooling times the heat capacity should be small and the heat conductivity of the strut should be large. It is also evident that the ratio T_h/T_c should be as nearly unity as possible. This requires that the tuner should produce the required displacement in the smallest temperature interval possible. If a given temperature difference is required to produce the necessary motion, then from a speed standpoint it is desirable to make both T_h and T_c large in order that their ratio shall be nearly unity. The allowable temperature is usually limited by constructional considerations.

Over the normal tuning range of a reflex oscillator we have previously shown that the tuning characteristic may be represented by

$$\lambda = \alpha \sqrt{C_f + C(x)} \tag{13.7}$$

where α is a constant

C_f is a lumped fixed capacitance

$C(x) = \dfrac{\beta}{x}$

β is a constant.

Hence one can show that for small changes Δx from x_0

$$\Delta\lambda = -\frac{\frac{1}{2}\lambda_0 \beta \Delta x}{x_0^2 C_0} \tag{13.8}$$

$$C_0 = C_f + C(x_0). \tag{13.9}$$

One may also show that for the type of tuner employed and the small motions involved in the 2K50 the displacement of the grids as a function of the temperature difference $T - T_0$ will be

$$x = x_0 - H(T - T_0) \tag{13.10}$$

whence

$$\Delta\lambda = \frac{\frac{1}{2}\lambda_0 \beta H (T - T_0)}{x_0^2 C_0}. \tag{13.11}$$

If at time $t = 0$, $x = x_0$, $T = T_0$, $\lambda = \lambda_0$, we have for heating

$$\lambda' - \lambda_0 = \frac{\frac{1}{2}\lambda_0 \beta H}{x_0^2 C_0}\left(\frac{P_i}{k} - T_0\right)(1 - e^{-(kt/c)}). \tag{13.12}$$

If we give P_i its maximum value P_m then at $t = \infty$ the temperature of the strut will be $\dfrac{P_i}{k} = T_m$, $\lambda = \lambda_m$

and $$\Delta\lambda = \lambda_m - \lambda_0 = \frac{\frac{1}{2}\lambda_0 \beta H}{x_0^2 C_0} (T_m - T_0) \tag{13.13}$$

or $$\lambda - \lambda_0 = (\lambda_m - \lambda_0)(1 - e^{-(kt/c)}). \tag{13.14}$$

Thus the behavior of this type of tuner may be described by a time constant which is given by $\tau = \frac{c}{k}$. This has been verified experimentally for the 2K50, in which this constant has been found to have a typical value of 1.3 seconds.

The instantaneous tuning rates at a given wavelength based on full on-full off operation can be shown to be

$$\frac{df}{dt} = -\frac{1}{\tau}\frac{f}{f_m}(f - f_m) \quad \text{heating} \tag{13.15}$$

$$\frac{df}{dt} = \frac{1}{\tau}\frac{f}{f_0}(f_0 - f) \quad \text{cooling} \tag{13.16}$$

where f_0 is the frequency at zero tuner power

f_m is the frequency at maximum permitted drive.

Figure 94 shows the instantaneous tuning rate as a function of frequency on heating and cooling.

Typical power output versus frequency characteristics for the 2K50 are shown in Fig. 95. Curve A shows the power output with the repeller voltage optimized at each frequency while Curve B gives the variation when the repeller voltage is set for an optimum at the center of the band and held fixed as the frequency changes. For constructional reasons the spacing between the repeller and second cavity grid is fixed in the 2K50 so that on a proportional frequency basis the range between half power points with fixed repeller voltage is smaller for the 2K50 than for the 2K45.

Figure 96 shows the frequency vs. grid voltage characteristics for the 2K50. For normal operation with full on-full off operation the grid voltage is switched between zero and cutoff.

H. *A Millimeter Range Oscillator*

During the latter stages of work on the 2K50 development, work was started on an oscillator for a wavelength range around .625 cm. The design of this developmental tube known as the 1464XQ was undertaken.

There are several difficulties in going from 1.25 cm. to .625 cm. Greater accuracy of construction is required and the cathode must be operated at a higher current density. The greatest difficulty arises from the fact that the grids cannot be directly scaled in size from those used in tubes for longer wavelengths.

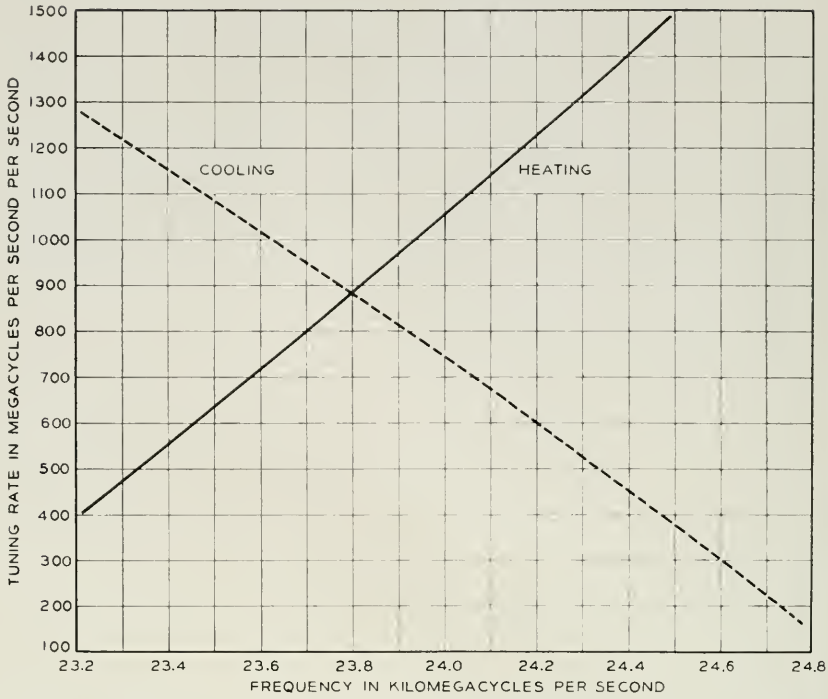


Fig. 94.—Computed instantaneous tuning rates on heating and cooling for the 2K50 oscillator. These results are based on a time constant of 1.3 seconds and on the assumption of "full on or off" operation.

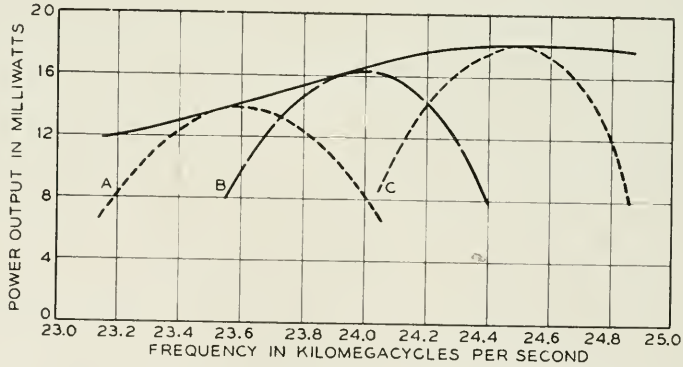


Fig. 95.—Power output characteristics of the 2K50 oscillator. Curves *A*, *B* and *C* illustrate the power variation with frequency when the repeller voltage is set for the optimum at the indicated frequency and held fixed as a cavity tuning is changed. The envelope of these curves shows the power variation with frequency when the repeller voltage is maintained at its optimum value for each frequency.

Let us consider the factors involved in scaling from a tube operating at a given frequency to a smaller tube operating at a higher frequency. If the cathode is operated space-charge-limited and the anode voltage is the same as for the larger tube, the total electron current will be the same and each grid wire will intercept as much current and hence receive the same power to dissipate as in the larger tube. Suppose the length of a grid wire in the larger tube is l_0 and in the smaller tube the length of the corresponding wire is l_1. Suppose that all the other dimensions of the smaller tube, including

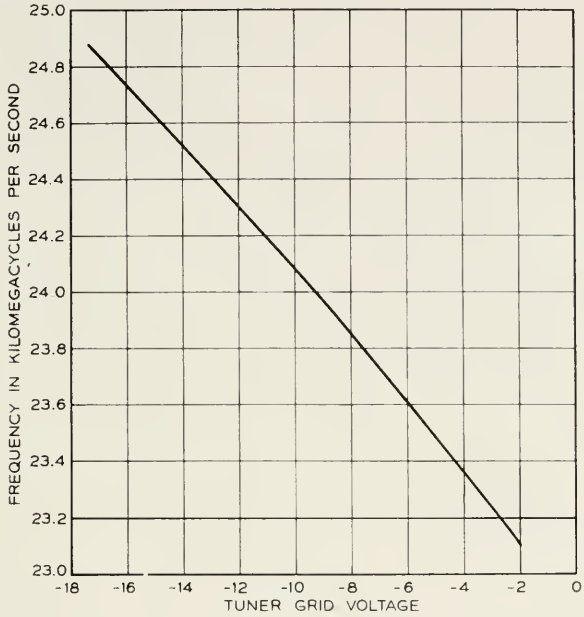


Fig. 96.—Frequency vs tuner grid voltage for the 2K50.

the diameter of the grid wire, are reduced in the ratio l_1/l_0. We do not know *a priori* whether or not the temperature distribution along the grid wires of the smaller tube will be the same as that for the larger tube; suppose, however, that it is. Then if T_0 is the temperature at some typical point, say, the hottest, on the grid wire of the larger tube, and T_1 is the temperature at the similar point on the smaller tube, the power the wire loses by radiation in the large tube, P_{r0} and in the small tube, P_{r1} are given by

$$P_{r0} = A l_0^2 T_0^4 \tag{13.17}$$

$$P_{r1} = A l_1^2 T_1^4. \tag{13.18}$$

Here A is nearly constant for given tube geometry and materials. In radiation the power lost varies as the area, which varies as l^2, and as the temperature to the fourth power.

The power lost by end cooling, for the large and the small tube, P_{e0} and P_{e1} will be given by

$$P_{e0} = Bl_0T_0 \tag{13.19}$$

$$P_{e1} = Bl_1T_1. \tag{13.20}$$

Here B is another constant. These relations express the fact that the power lost by end cooling (thermal conduction) varies as cross sectional area divided by length and hence as l and as temperature difference, taken as proportional to T.

Now, in scaling the tube the power to be dissipated has been kept constant. Further, in making the tube small, the hottest point of the grid cannot be run hotter than the melting point of the wire; in fact, it cannot be run nearly this hot without unreasonable evaporation of metal. Suppose we let the grid in the smaller tube attain the maximum allowable temperature T_m and let the power the wire must dissipate be P. Then for the large tube

$$P = P_{r0} + P_{e0} = (Al_0T_0^3 + B)l_0T_0 \tag{13.21}$$

and for the smaller tube

$$P = P_{r1} + P_{e1} = (Al_1T_m^3 + B)l_1T_m. \tag{13.22}$$

Hence, the smallest value l_1 can have without running the grid too hot is given by the equation

$$l_1 = l_0 \frac{T_0}{T_m} \frac{(Al_0T_0^3 + B)}{(Al_1T_m^3 + B)}. \tag{13.23}$$

We see that if l_0 is very small,

$$\begin{aligned} Al_0T_m^3 &\ll B \\ Al_1T_m^3 &\ll B \end{aligned} \tag{13.24}$$

Numerical examples show that this is so for a tube such as the 2K50. This means that nearly all of the power dissipated by the grid is lost through end cooling, not radiation.[14] Further, in the 2K50 the grid is already operating near the maximum allowable temperature. Hence, nearly, $T_0 = T_m$ and the smallest ratio in which the tube can be scaled down without overheating the grids is approximately unity. This means that in making a tube for .625 cm. the grid wire cannot be made half the diameter of the wire used in

[14] The fact that one kind of dissipation predominates in both cases justifies the assumption of the same temperature distribution in both cases.

the 2K50. In fact, the diameter of the grid wire can be made but little smaller. Thus, in the 1464XQ the grid is relatively coarse compared with that in the 2K50. This results in a reduced modulation coefficient and hence in less efficiency.

In going to .625 cm., resonator losses are of course greater. The surface resistivity of the resonator material varies as the square root of the frequency. Surface roughness becomes increasingly important in increasing resistance at higher frequencies. Further, in order to provide means for moving the diaphragm in tuning, it was necessary in the 1464XQ to use a second mode resonator (described later) and this also increases losses over those encountered at lower frequencies.

Development of the 1464XQ was stopped short of completion with the cessation of hostilities. However, oscillation in the range .625–.660 cm. had been obtained. The power output varied from 2–5 milliwatts between the short wave and long wave extremes of the tuning range. The cathode current was around 20 ma, the resonator voltage 400 volts. The tube operated in several repeller modes in a repeller voltage range 0 to −180 volts.

Figures 97 and 98 illustrate features of the 1464XQ oscillator. Figure 98 is a scale drawing of the resonator and repeller structure. The electron beam is shot through two apertures covered with grids of .6 mil tungsten wire. These grids are 80% open and are lined up. The aperture in the grid nearest the gun is 23 mils in diameter and the second aperture is 34 mils in diameter. The repeller is scaled almost exactly from the 723A 3 cm. reflex oscillator. A second mode resonator is used. The inner part, a, of Fig. 98 is about the size of a first mode resonator. This is connected to an outer portion, c, by a quarter wave section of small height, b, which acts as a decoupling choke. The resonator is tuned by moving the upper disk with respect to the lower part, thus changing the separation of the grids. The repeller is held fixed. Power is derived from the outer part, c, of the resonator by means of an iris and a wave guide, which may be seen in the section photograph Fig. 97. There is an internal choke attached to the end of the part of the wave guide leading from the resonator. This is opposed to a short section of wave guide connected to the envelope, and in the outer end of this wave guide there is a steatite and glass window of a thickness to give least reflection of power.

I. *Oscillators for Pulsed Applications—The 2K23 and 2K54*

All the reflex oscillators described in the preceding sections have been low power oscillators intended for beating oscillator or signal oscillator applications. Some limitation on the power capability of these oscillators in the form previously described is set by the power handling capacity of the grids. If the tubes are pulsed with pulse durations which are short compared with

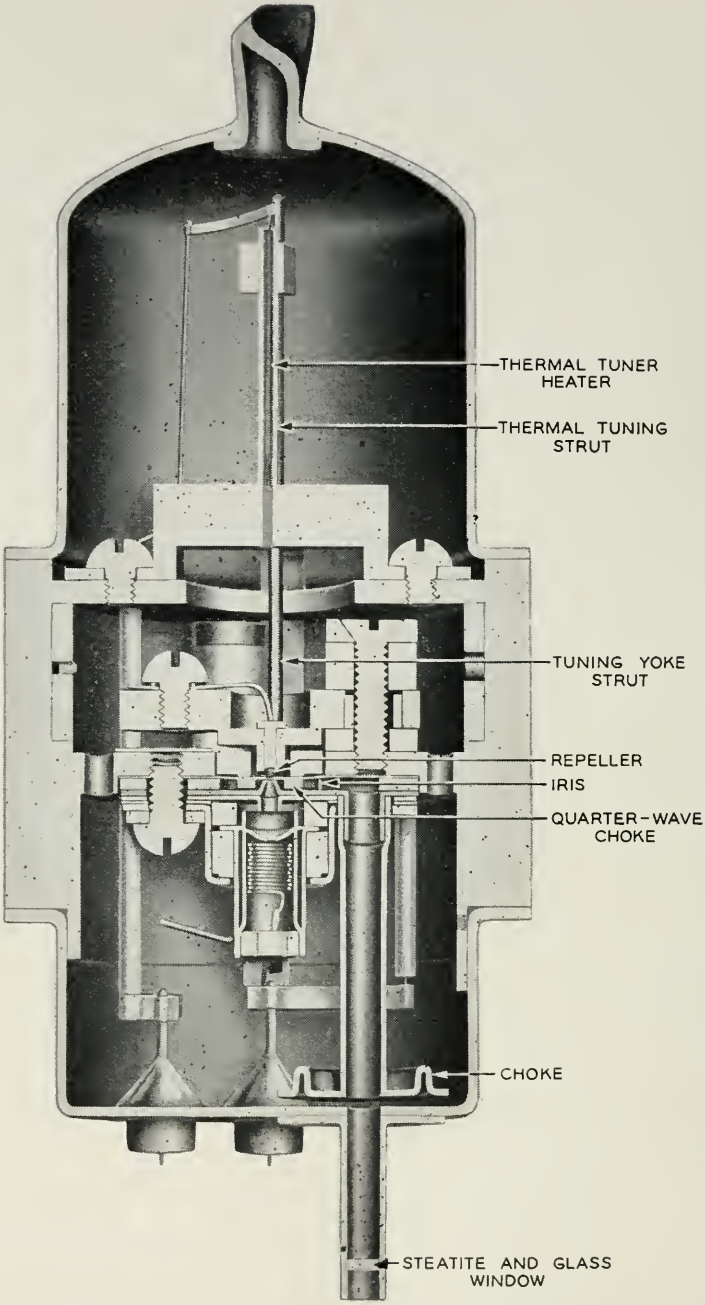


Fig. 97.—An experimental thermally tuned reflex oscillator, the 1464, designed for operation between the wave lengths of .6 and .7 cms.

the thermal time constant of the grids, then the peak power input to the oscillators may be increased over the continuous limit by the duty factor, provided that the voltages applied are consistent with the insulation limits of the tubes and that the peak currents drawn from the cathode are not in excess of its capacity.

An application for a pulsed oscillator arose in the AN/TRC-6 radio system.[15] This was an ultra-high frequency military communication system using pulse position modulation to convey intelligence. With the high gain which may be achieved with antennas in the centimeter range, the power necessary in the transmitter for transmission over paths limited by line of sight is of the order of a few watts peak. In beating oscillator applications the power output is of secondary importance to electronic tuning, so that the reflex oscillators previously described were designed to operate with a drift time in the repeller space such as to provide the desired tuning. In a

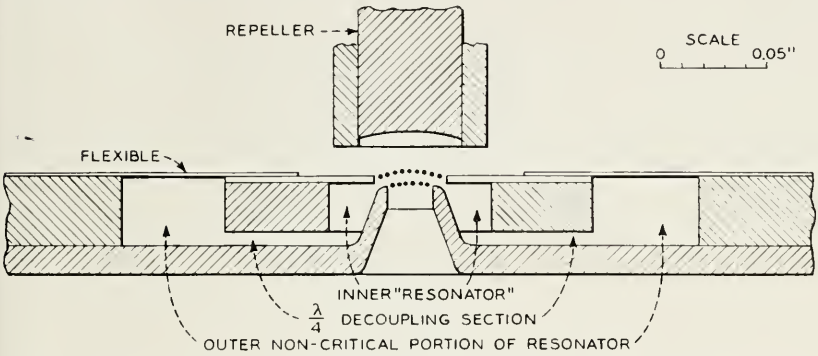


Fig. 98.—The resonator and repeller structures of the oscillator shown in Fig. 97.

pulsed transmitter electronic tuning is unnecessary and indeed undesirable, since it leads to frequency modulation on the rise and fall of the pulse. In section III it is shown that for maximum efficiency with a given resonator loss there will be an optimum value for the drift time. If there were no resonator loss this time would be $\frac{3}{4}$ cycles, which is the minimum possible. By utilizing the optimum drift angle and taking advantage of the higher peak power inputs permitted by pulse oscillator it was possible to obtain peak power outputs of the order of several watts using the same structure as employed for the beating oscillators previously described without exceeding the power dissipating capability of such a structure. From the standpoint of military convenience this was a very desirable situation for reasons of simplicity of tuning and ease of installation.

[15] A Multi-channel Micro-wave Radio Relay System, H. S. Black, W. Beyer, T. J. Grieser, and F. A. Polkinghorn, *Electrical Engineering* Vol. 65, No. 12, pp. 798–806, Dec. 1946.

The first tube designed for this service was the 2K23. It was based on the design employed in the 2K29 and operated with a drift time of $1\frac{3}{4}$ cycles in the repeller space. A severe limitation on the performance was set by by the requirement that a single oscillator should cover the frequency range from 4275 to 4875 Mc/s. It is shown in Section X that in an oscillator tuned by changing the capacitance of the gap the efficiency will vary considerably

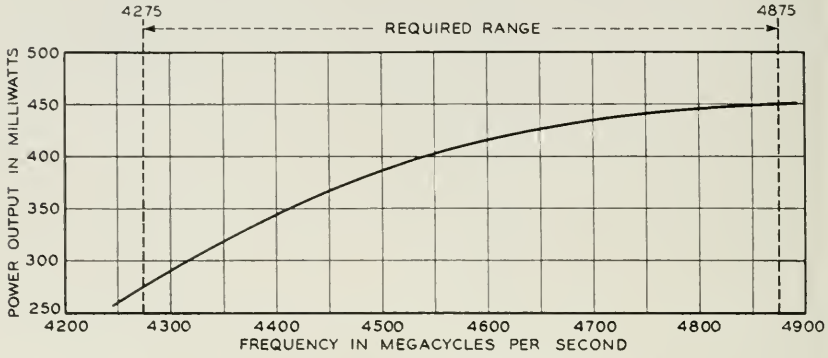


Fig. 99.—Variation of the peak power output vs frequency for the 2K23 reflex oscillator. This tube was designed for pulse operation with a duty factor of 10 in a repeller mode having 3.5 π radians drift.

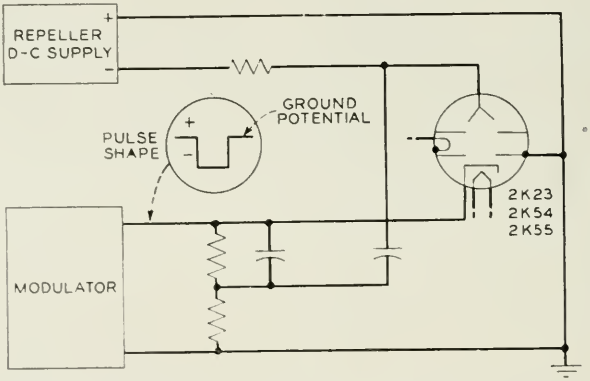


Fig. 100.—Modulator circuit for use in connection with reflex oscillators showing means for applying pulse voltage to the repeller to reduce frequency modulation during pulsing.

over such a tuning range. In the case of the 2K23 the variation of the peak power output with frequency is shown in Fig. 99. The duty factor at which the tube was used (the ratio of the time between pulses to the pulse length) was 10. In the AN/TRC-6 application the tube was operated on a fixed current basis; i.e., the pulse amplitude was adjusted to a value such that the average current drawn was 15 ma. The schematic of the circuit employed in pulsing oscillators in this way is shown in Fig. 100. The resonator

of the oscillator is operated at ground and the cathode of the oscillator is pulsed negative with respect to this ground. The repeller voltage is referenced from the cathode and this reference is maintained during the pulse. Some frequency modulation occurs during the rise and fall of the pulse because of the changing electron velocity. This can be reduced in part by applying a part of the pulse in the repeller circuit proportioned in such a way as to tend to maintain the drift time independent of the cathode to resonator voltage. Satisfactory performance is achieved in this way.

As mentioned previously, the intelligence is conveyed by pulse position modulation. The AN/TRC-6 system uses time division multiplex to provide eight communication channels. The multiplexing is achieved by transmitting a four micro-second marker pulse which provides a time reference followed by eight one micro-second pulses. The time of each of the latter pulses is independently varied in position with reference to the marker pulse.

The time interval from the marker to each pulse could be measured to either the leading or trailing edge of the pulse. In Section XII it is shown that the leading edge of the r.f. pulse will be subject to what is commonly called "jitter" because of the random time of rise which will result if oscillation starts from shot or Johnson noise. Conceivably oscillation might be started by shock excitation of the resonant circuit by the pulsed beam current. However, in Appendix X it is shown that the initial excitation produced by shot noise in the beam exceeds that induced by the current transient by a factor of approximately 100. The trailing edge of the pulse will not be subject to this form of jitter provided two conditions are met. First, the pulse duration must be long enough so that oscillation builds up to full amplitude during the pulse. Second, the receiver must have a sufficient bandwidth so that the transient which occurs on reception of the leading edge has fallen to a small value by the time the trailing edge is received. Since these conditions were met in the AN/TRC-6 system, the trailing edge of the pulse was used.

In the latter stages of the development of the AN/TRC-6 system it was decided to remove the restriction that the required tuning range should be covered with a single transmitter tube. This made possible the achievement of a design which would provide an improved performance throughout the band. In order to improve the circuit efficiency of the resonator, the new designs were based on the oscillator structure which was employed in the 2K45. It has been pointed out previously that this makes possible a considerably higher resonant impedance of the cavity, partly because of the reduced gap capacitance and also because the smaller first and second grids reduce the resonator losses. These effects were reflected in the higher efficiency obtained in the 2K54 and 2K55.

The three tubes, the 2K23, 2K54 and 2K55 were intended to couple to wave guide circuits. Although direct coupling to the guide through a wave guide output would have been desirable in some ways, it would have resulted in a very large and awkward structure. It seemed best, therefore, to retain the coaxial output feature of tubes designed for beating oscillator service. From a standpoint of convenience, it would have been desirable

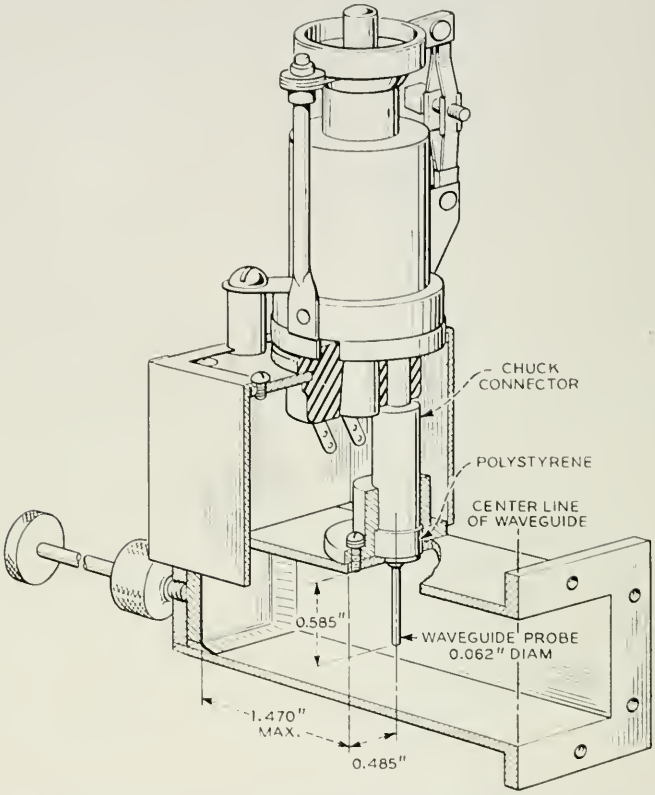


Fig. 101.—A transducer designed to adapt the 2K23, 2K54 and 2K55 oscillators to a terminated wave guide load. If the back piston of this coupling is set at a distance of 1.08" from the center of the probe the impedance presented to the oscillator line is 50 ohms with a 1 db variation over the required frequency range.

to have had the wave guide probe an integral part of the tube, as is the case in the 2K25. However, the length of a quarter wave probe at the operating wave lengths of the AN/TRC-6 system made such a design hazardous. Because of this a transducer was developed into which the tube could be plugged. This is shown in Fig. 101. The central chuck extends the center conductor of the output coaxial of the tube into the guide. It had been

originally intended to design this transducer so that all elements would be fixed. It was found, however, that an adjustable back piston setting permitted compensation for manufacturing tolerances in the tubes as well as permitting the presentation of a better impedance to the oscillator over the band than was attainable with a fixed transducer. This is illustrated in Fig. 102 which shows the power output versus frequency for a typical 2K54 and a typical 2K55 as a function of frequency for three cases. In one case, as shown by curve B, the power output is delivered at all frequencies into the optimum impedance, i.e. the impedance into which the oscillator will deliver maximum power. Curve *A* shows the power output delivered into

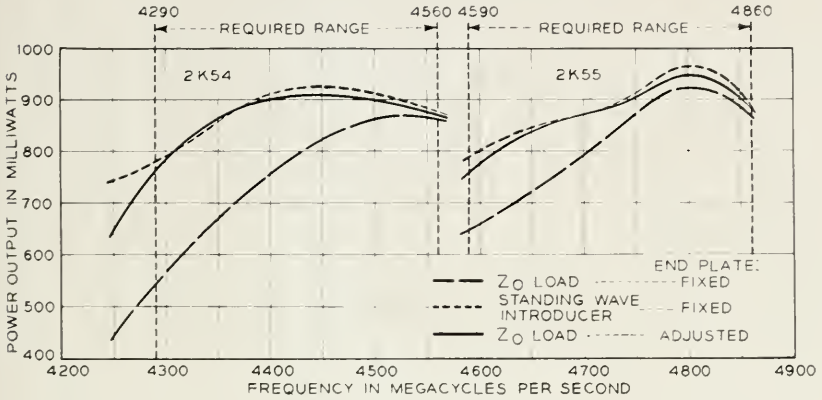


Fig. 102.—Peak power output vs frequency for the 2K54 and 2K55 oscillators for several load conditions. Curves *A* give the performance obtained when the tubes operate into a characteristic admittance load through the coupling of Fig. 101 with the end plate fixed so that the admittance presented to the tubes is constant to within 1 db over the frequency range. Curves *B* give the performance obtained when the optimum impedance is presented to the oscillators throughout the band. Curves *C* give the performance obtained when the tubes are coupled to a characteristic admittance load with the coupling unit of Fig. 101 and with a back piston adjusted to the best value at each frequency.

a transducer which has the back piston fixed at a distance from the probe, so chosen that the impedance seen by the oscillator is flat over the band to within 1 *db*. Curve *C* shows the power output when the back piston of the transducer is adjusted so that the tube delivers maximum power. It can be seen that the performance obtained under the latter circumstances is very nearly as good as that obtained when the optimum impedance is presented to the oscillator.

From Sections III and IX we would expect that, since the coupling system is such as to give maximum power throughout the band, the sink margin should be slightly greater than 2. Figures 103 to 106 give impedance performance diagrams for 2K54 and 2K55 at the four transmitting frequencies

of the AN/TRC-6 system. From this it can be seen that the sink margin is approximately 2 at all frequencies. In a transmitter tube another factor, which is of small importance in a beating oscillator, becomes of interest. This factor is the pulling figure, which is defined as the maximum frequency

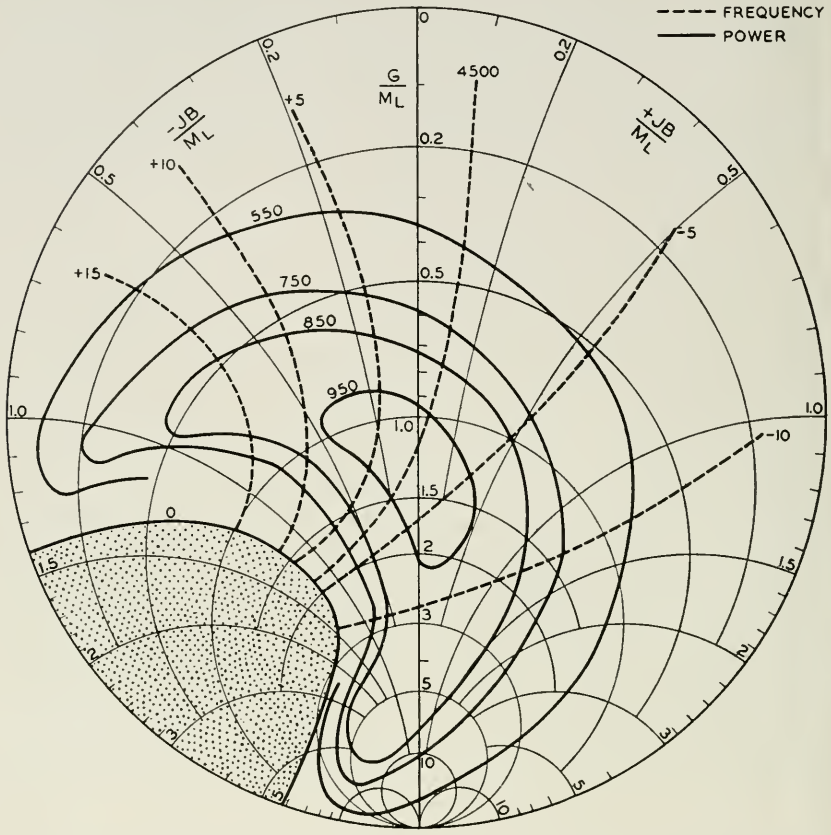


Fig. 103.—Rieke diagram for the 2K54 at a nominal frequency of 4500 megacycles. The point at the unity *vswr* condition is obtained by adjusting the repeller voltage and the back piston of the coupling of Fig. 101 to the values which gave maximum power. These conditions were then held fixed for the remainder of the chart.

excursion which will be produced when a VSWR of 3 *db* is presented to the transmitter and the phase is varied over 180°. Fig. 107 gives the pulling figure as a function of frequency for the 2K54 and 2K55. The requirements on the pulling figure for the 2K54 and 2K55 were not severe, since in the AN/TRC-6 system the tube is coupled to the antenna by a very short wave guide run and, furthermore, the antennas are fixed.

Investigation of the pulling figure of early models of the 2K55 led, however, to the disclosure of one unforeseen pitfall arising from the existence of electronic hysteresis. It had at first been considered that electronic hysteresis would not be of importance in the transmitter tube, where the feature of electronic tuning was of no importance. This might be true in a CW oscilla-

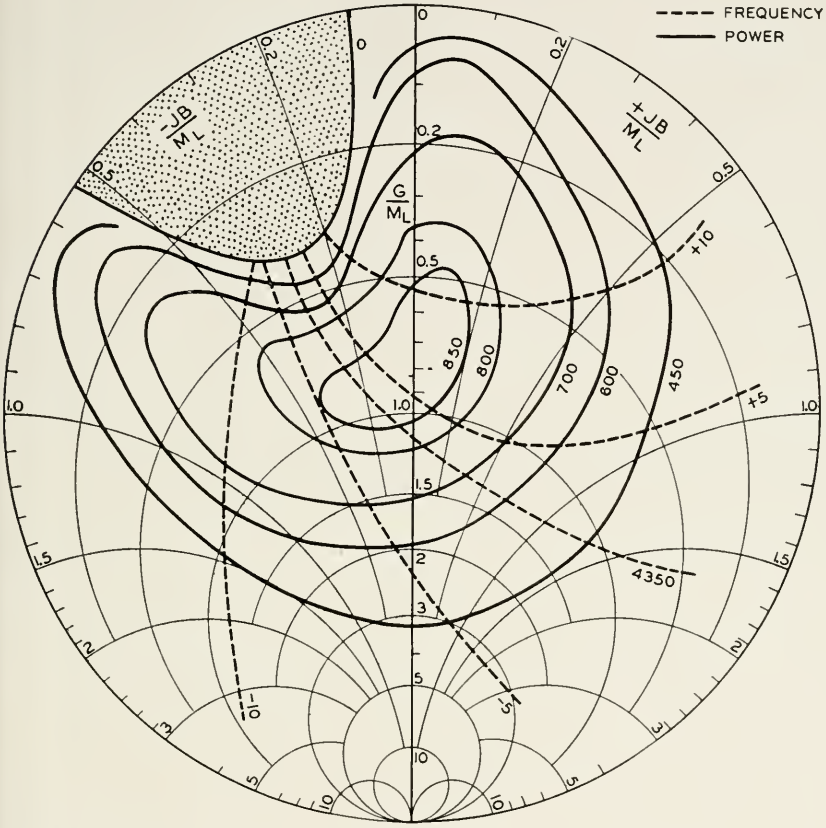


Fig. 104.—Rieke diagram for the 2K54 oscillator at a nominal frequency of 4350 megacycles. The unity *vswr* point was obtained as described in Fig. 103.

tor, but in a pulsed oscillator the existence of hysteresis resulted in an unforeseen reduction of the sink margin. Since the oscillator is being pulsed, for each pulse the oscillating conditions are being re-established. Although the cathode-repeller voltage need not vary during the pulsing, the fact that the cathode-resonator voltage is being changed means that for each pulse the drift angle in the repeller space varies on the rise and fall of the pulse. The effect during the rise of the pulse is the same as though the repeller

voltage were made less negative. In other words, on the rise of each pulse the situation is equivalent to that in a CW oscillator when, for a fixed resonator voltage, one starts with a repeller voltage too negative to permit oscillation and then reduces the repeller voltage until oscillation occurs. Let us now suppose that the hysteresis is such that under these circumstances

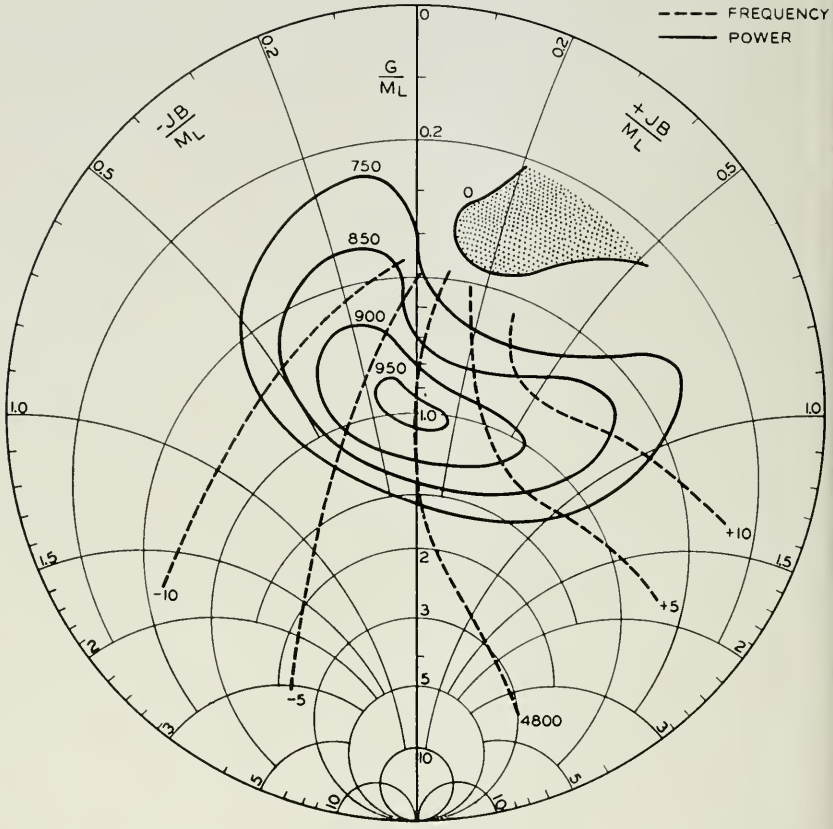


Fig. 105.— Rieke diagram for the 2K55 oscillator at a nominal frequency of 4800 megacycles. The unity *vswr* point was obtained as described in Fig. 103.

the amplitude of oscillation would suddenly jump to a large value. We would then obtain a variation of peak power output as a function of the repeller voltage shown in Fig. 108. Ordinarily, the repeller voltage would be adjusted so as to obtain maxmum power output as, for example, with the repeller voltage V_{R1}. Next, let us suppose that a variable impedance is presented to the oscillator with the repeller voltage held fixed at value V_{R1}, as would be done, for example, in obtaining an impedance performance

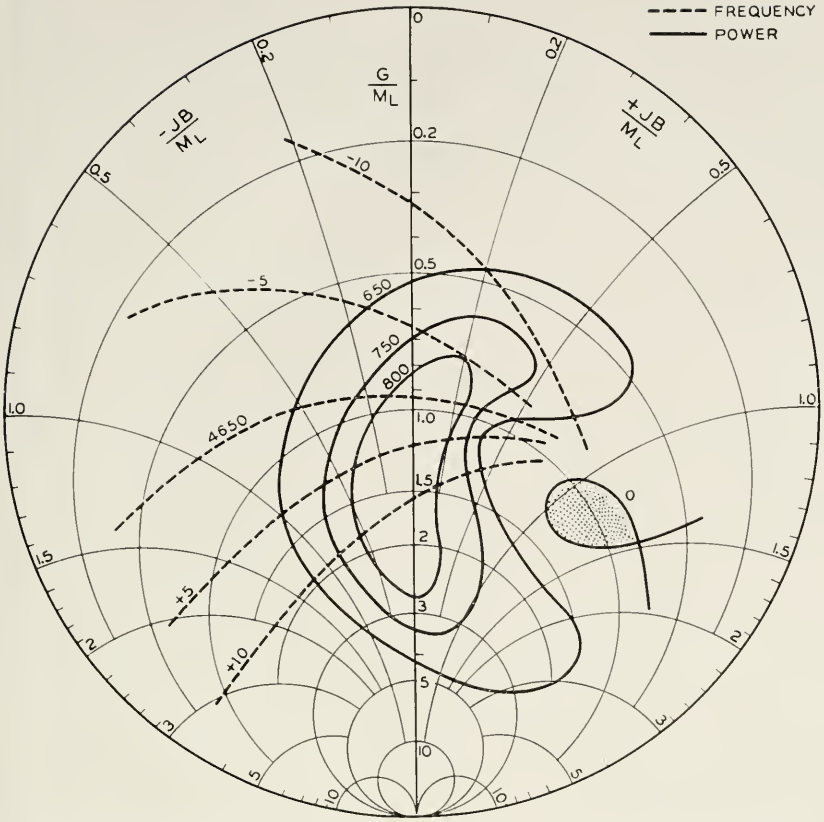


Fig. 106.—Rieke diagram for the 2K55 oscillator at a nominal frequency of 4650 megacycles. The unity *vswr* point was obtained as described in Fig. 103.

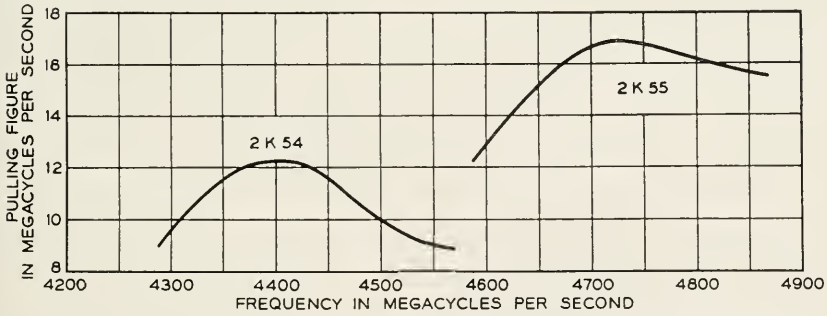


Fig. 107.—Pulling figures for the 2K54 and 2K55 reflex oscillators as functions of frequency. With a unity *vswr* load the repeller voltage and back piston of the coupling of Fig. 101 were adjusted for a maximum power and held fixed for the pulling figure measurements.

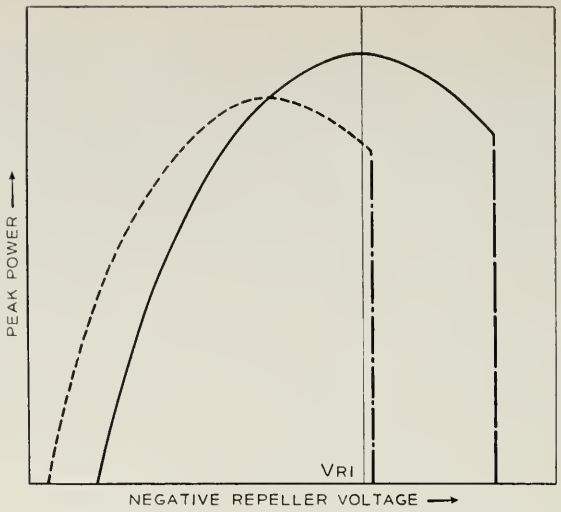


Fig. 108.—The variation of power output with repeller voltage for a pulsed re 'ex oscillator exhibiting hysteresis.

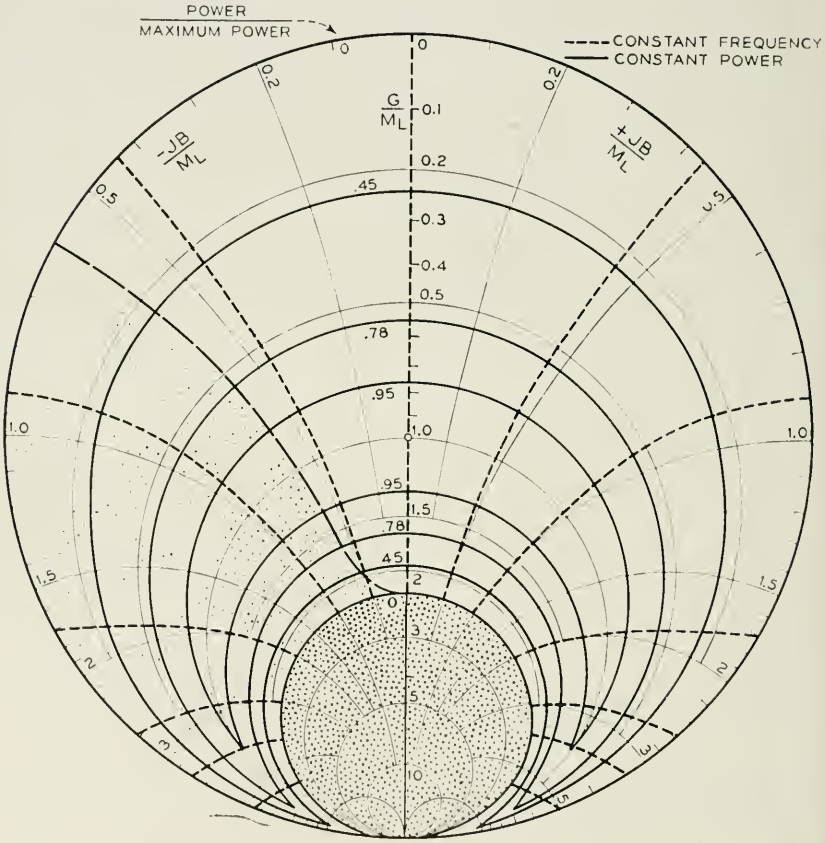


Fig. 109.—The effect of hysteresis on the Rieke diagram of a pulsed reflex oscillator. The hysteresis shown in Fig. 108 can result in failure of a pulsed oscillator to operate in the lightly shaded portion of the Rieke diagram as well as the heavily shaded portion corresponding to the normal sink.

diagram. The effect of varying the impedance on the repeller characteristic in Fig. 108 is to shift the whole characteristic to the right or left, depending upon the phase of the impedance, as well as to change its general form as shown by the dotted curve. It can be seen from this that if the hysteresis is sufficiently bad and if the pulling figure exceeds a particular value, one

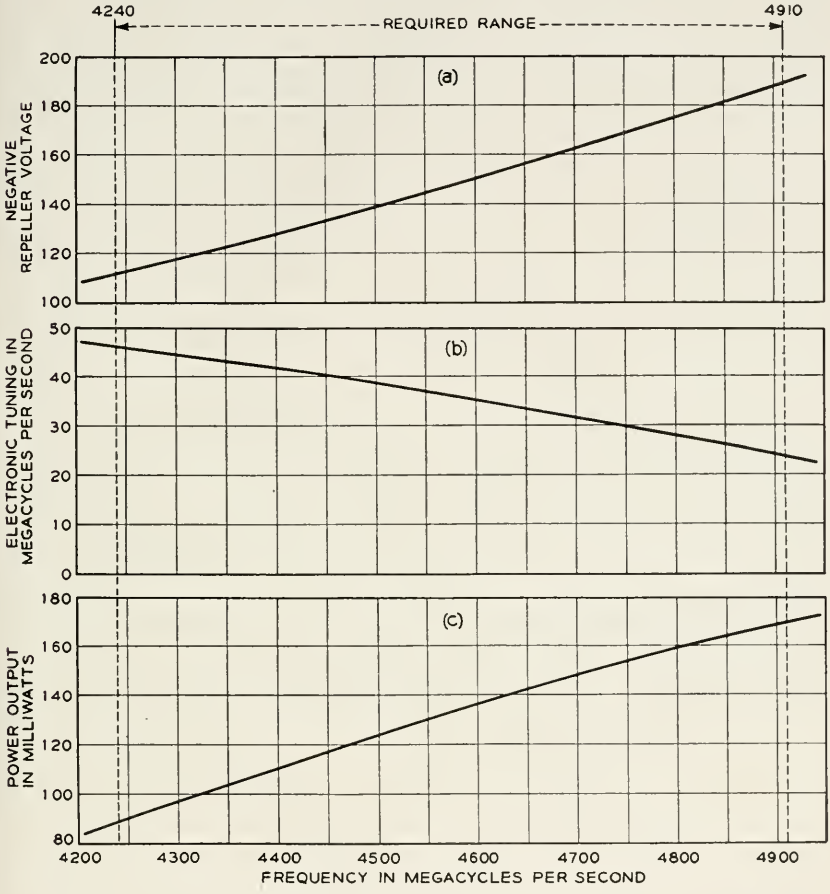


Fig. 110.—Performance characteristics of the 2K22 operated into a 50 ohm load.

effect of the hysteresis will be to reduce the sink margin, and this is found to be the case. Fig. 109 shows an impedance performance diagram obtained with pulsed operation for an early model of the 2K55 in which the hysteresis was excessive. From this it can be seen that the area of the sink on the diagram is very greatly increased and also that the sink margin is reduced from the theoretical value of somewhat in excess of 2 to a value of less than

1.1 although maximum power occurs at unity VSWR. A modification in the design reduced the hysteresis and eliminated this effect as shown by Figs. 103 to 106.

In addition to the transmitter tube for the AN/TRC-6 system, it was necessary to design a beating oscillator. This tube is known as the Western Electric 2K22. Its design was scaled from the 2K29, previously described. The tube was designed to operate into a 50 ohm impedance and can be coupled by a coaxial adapter to a 50 ohm line or by means of the transducer of Fig. 101 to a wave guide. When the back piston of the transducer is set at a distance of 1.080″ from the probe center, the impedance presented to the oscillator is 50 ohms with approximately a 1 *db* variation over the frequency range. Fig. 110 gives the performance characteristics of the 2K22 operating into a 50 ohm load. The AN/TRC-6 system using these tubes provided a military communication system during the war. A description of this service has been given.[12] Also, models of the AN/TRC-6 system have been put into service to provide telephone communication between Cape Cod and Nantucket and also between San Francisco and Catalina Island.

J. *Scope of Oscillator Development at the Bell Telephone Laboratories*

The reflex oscillators discussed in the foregoing sections were developed primarily for beating oscillator service, and in one instance for a transmitter. Reflex oscillators also received wide application in test equipment. The best-known application of this type was in the spectrum analyzer, in which the electronic tuning characteristic of the oscillators made possible the display of output spectra, and especially of the spectra of magnetron oscillators, on an oscilloscope. This greatly facilitated the development of the magnetron. The reflex oscillator also was widely used as a signal generator, and the ease of frequency adjustment particularly suited it to this application. In some signal generators it was desired to pulse the reflex oscillator at low power levels. As an alternative to the method previously described, in which the voltage between the cathode and resonator was pulsed, it is possible to leave this voltage fixed and to pulse the voltage between the repeller and cathode. In this case the repeller-cathode voltage is set at a base value at which the tube will not oscillate and the pulse varies this voltage to the oscillating value.

The oscillators which have been described have been chosen to indicate various features of their development. In addition to these, a number of other oscillators were developed to meet various service needs. Figure 111 shows a chart giving the frequency ranges of these tubes. Of the eleven beating oscillators of the reflex oscillator type on the Army-Navy preferred

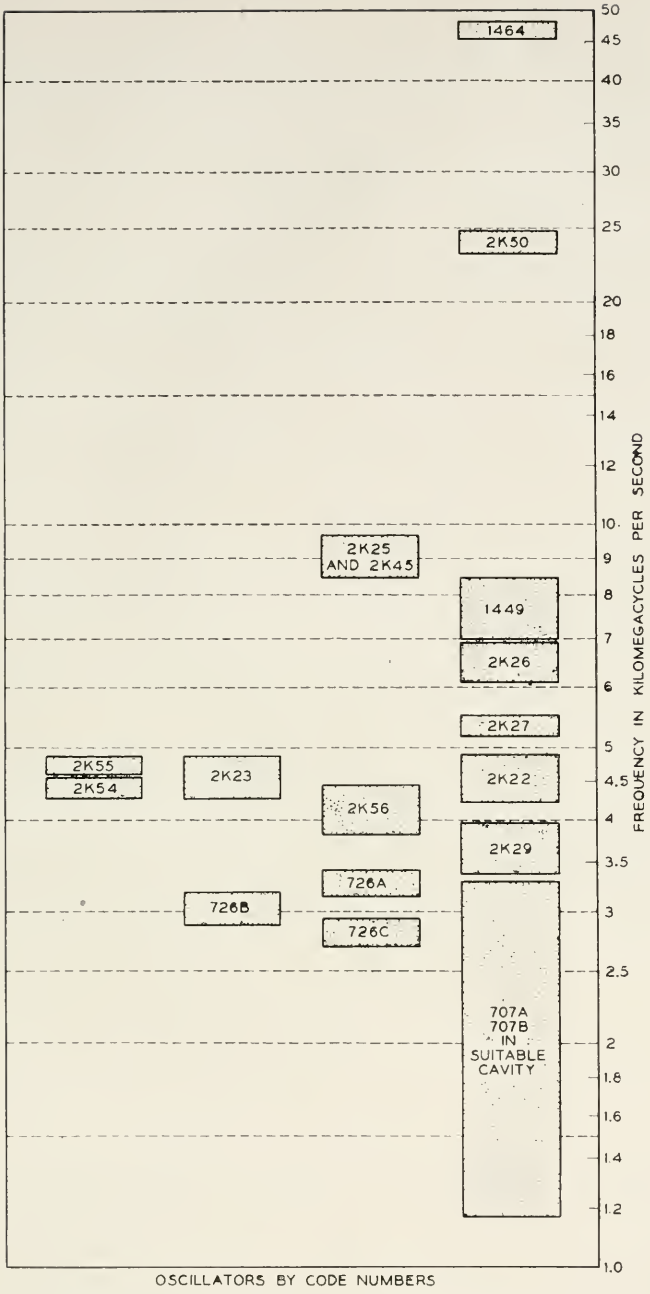


Fig. 111.—Reflex oscillators developed at the Bell Telephone Laboratories. The boxes around the tube numbers show the frequency range covered.

list of electron tubes for 1945, nine were developed at the Bell Telephone Laboratories.

APPENDIX I

Resonators

In thinking about resonators it is important in order to avoid confusion to keep a few fundamental ideas in mind. One of the most important is that we must not use the notion of scalar potential in connection with fluctuating magnetic fields. Electric fields produced by fluctuating magnetic fields cannot be derived from a scalar potential, and in the presence of such fields to speak about the potential at a point is hopelessly confusing.[16]

The idea of voltage as the line integral of electric field along a given path between two points is useful, but it must be remembered that the voltage depends on the path chosen. Consider, for instance, an ordinary 60-cycle transformer with the secondary wound of copper tubing. For a path from one secondary terminal to the other through the center of the tubing the voltage (integral of field times distance) is zero. For a path between terminals outside of core and coil, the voltage between terminals is $d\psi/dt$, where ψ is the magnetic flux linkage of the path and the coil, counting each line of force as many times as the path encircles it.

If resistance drop is neglected the work done in moving a charge through a conductor is zero. The line integral of an electric field around a closed path is $d\psi/dt$. If part of the path is through a conductor, or through a space where there is no electric field, the voltage along the rest of the path (as between portions of the conductor) is $d\psi/dt$. For paths linking different amounts of flux, the voltage will be different. In the case of low frequency transformers, all paths linking the terminals and lying outside of the core and coil link practically the same amount of flux, and there is little ambiguity about the voltage. In reflex oscillators the electrons travel from one field free region to another along a certain path and this determines the path along which the voltage should be evaluated.

To review: the voltage between two points is the integral of the field along the path times distance, and refers to a certain path. If the path begins and ends in a field free region, the voltage is $d\psi/dt$, where ψ is the magnetic flux linking the chosen path and a return path through the field free region.

To this should be added that high frequency currents and fields penetrate the surface of metals only a fraction of a thousandth of an inch in the centimeter range,[17] so that the interior of a conductor is field free, and fields inside of a metal enclosed space cannot produce fields outside of that space.

[16] The electric field can, of course, be derived from a scalar and a vector potential.

[17] As an example, for copper the field is reduced to (1/2.72) of its value at the surface

In considering resonators we should further note that the magnetic flux must be produced by a flow of current, either convection current or displacement current, around the lines of force. In a transformer this can be identified as the current flowing in the coils. In the resonators used in reflex oscillators it is the current flowing in the walls and, as displacement current, across the gap and from one face of the resonator to the other.

Two axially symmetrical resonators suitable for use in reflex oscillators are shown schematically in Figs. 112 and 113. The resonator in Fig. 112 has grids and might be used with a broad unfocused electron beam at a low d-c voltage; that shown in 113 has open apertures and might be used with a

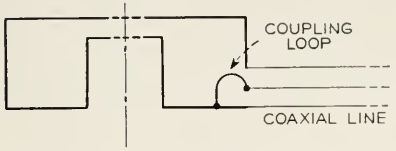


Fig. 112.—An oscillator cavity with grids and loop coupling to a coaxial line.

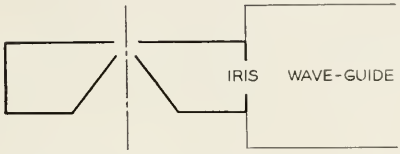


Fig. 113.—An oscillator cavity without grids and with iris coupling to a wave guide

focused high voltage electron beam. In Fig. 112, the resonator is coupled to a coaxial line by means of a coupling loop or coil; in Fig. 113 the resonator is coupled to a wave guide by means of a small aperture or "iris."

Let us consider the resonator of Fig. 112 in the light of what we have just said. A magnetic field flows around the axis inside of the resonator. There is an electric field between the top and bottom inside surfaces of the resona-

---

at a depth.

$$\delta = 3.82 \times 10^{-5}\sqrt{\lambda} \qquad (a1)$$

Here λ is wavelength in centimeters. It may also be convenient to note that the surface resistivity of a centimeter square of resonator surface is, for copper,

$$R = .045/\sqrt{\lambda} \qquad (a2)$$

This means that if a current of I amperes flows on a surface over a width W and a length l, the power dissipation is

$$P = I^2 R l / W \qquad (a3)$$

For other non-magnetic metals, both δ and R are proportional to the square root of the resistivity with respect to that of copper.

tor. There is no electric field outside of the resonator (except a very little that leaks out near the grids). To take a charge from one grid to another *outside* of the resonator requires no work. Thus the voltage across the gap is $d\psi/dt$, where ψ is the magnetic flux around the axis. Actually, there is a little magnetic flux between the grids, and hence the voltage near the edges of the grids is a little less than that at the center.

Current flows as a displacement current between the grids, and as convection current radially out around, and back along the inside of the resonator to the other grid. This current flow produces the magnetic field that links the axis.

Part of the magnetic flux links the coupling loop. If this part is ψ_ℓ and if the coupling loop is open-circuited, the voltage across the coupling loop will be $d\psi_\ell/dt$.

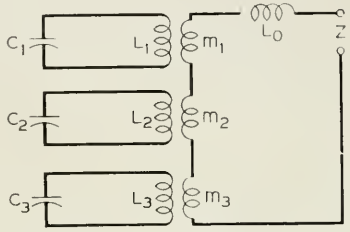


Fig. 114.—Equivalent circuit for a resonator having 3 modes of resonance.

In the resonator of Fig. 113, power leaks out through the iris into the wave guide. Part of the wall current in the resonator flows out through the hole into the guide; part of the magnetic flux in the resonator leaks out into the guide.

In dealing with resonators as resonant circuits of reflex oscillators, to which the electron stream and the load are coupled, we are interested in the gap and output impedances. For a clear and exact treatment, the reader is referred to a paper by Schelkunoff.[18] No exhaustive treatment of the problem will be given here, but a few important general results will be given.

If the resonator is lossless, the impedance looking into the loop may be represented exactly by an equivalent circuit indicated in Fig. 114. As coils used at low frequencies are really not simply ideal "inductances," (an idealized concept), but have many resonances (ascribed to distributed capacitance), so the resonator has many, in fact, an infinity of resonances. In the equivalent circuit shown in Fig. 114, only 3 of these are represented,

[18] S. A. Schelkunoff, Representation of Impedance Functions in Terms of Resonant Frequencies, *Proc. I.R.E.*, 32, 2, pp. 83–90, Feb., 1944.

by the inductances L_1, L_2, L_3 and the capacitances C_1, C_2, C_3. These resonant circuits are coupled to the terminals by mutual inductances m_1, m_2, m_3. In series with these appears the inductance measured at very low frequencies L_0, the self inductance of the coupling loop. The circuit in Fig. 114 may be regarded as a symbolic representation to be used in evaluating Z, just as a mathematical expression may be a symbolic representation of the value of an impedance.

In practical cases, the resonances are usually considerably separated in frequency, and near a desired resonance the effect of others may be neglected. In addition, if the Q is high we may add a conductance G_R across the capacitance to represent resonator losses (Fig. 115). It would be equally legitimate to add a resistance in series with L. In Fig. 115 a load impedance Z_L has been added. Fig. 115 is a very accurate representation of a slightly lossy resonator, a low loss coupling loop, and a load impedance. The

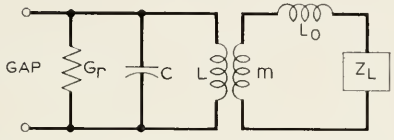


Fig. 115.—Equivalent circuit showing connection between the oscillator gap regarded as one pair of terminals and the oscillator load for an oscillator resonator having only one resonant frequency near the frequency of operation.

meaning of L and C will be made clearer a little later. We will now clarify the meaning of m. Suppose no current flows in the coupling loop ($Z = \infty$). Let the peak gap voltage be V. The peak voltage across m will be

$$V_m = mV/L \tag{a4}$$

In a resonator, if a peak voltage V across the gap produces a peak flux ψ_m linking the coupling loop when no current flows in the coupling loop, then

$$V_m = d\psi_m/dt = mV/L \tag{a5}$$

This defines m in terms of magnetic field, and L.

Figure 115 is also a quite accurate representation of Fig. 113. In this case the "terminals" are taken as located at the end of the wave guide. L_0 is the inductance of the iris, which will vary with frequency.

When we are interested in the impedance at the gap as a function of frequency, we may equally well use the equivalent circuit of Fig. 116. Here G_L represents conductance due to load; G_R represents conductance due to resonator loss. The total conductance, called G_C, is

$$G_C = G_R + G_L \tag{a6}$$

The circuit of Fig. 116 does not tell us how G_L varies with variation in load impedance Z_L. Further, L and C in this circuit are changed in changing Z_L, and include a contribution from the inductance of the coupling loop (which should not be very large).

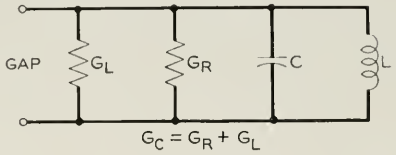


Fig. 116.—A simplified equivalent circuit of an oscillator resonator.

Several resonator parameters are vitally important in discussing reflex oscillators. These will be discussed referring to Fig. 116. G_L and G_R have already been defined. The resonant radian frequency ω_0 is of course

$$\omega_0 = (LC)^{-1/2} \tag{a7}$$

A very important quantity will be called the characteristic admittance M

$$M = (C/L)^{1/2} \tag{a8}$$

This quantity is important because at a frequency $\Delta\omega$ off resonance the admittance of the circuit is very nearly

$$\begin{aligned} Y &= G + j2M\Delta\omega/\omega_0 \\ \Delta\omega &= \omega - \omega_0 \end{aligned} \tag{a9}$$

The "loaded" Q of the circuit will be referred to merely as Q and is

$$Q = M/G_C \tag{a10}$$

The unloaded Q is

$$Q_0 = M/G_R \tag{a11}$$

A quantity which may be called the "external Q" is

$$Q_E = M/G_L \tag{a12}$$

We see

$$1/Q_0 + 1/Q_E = 1/Q \tag{a13}$$

The energy stored in the magnetic field at zero voltage across the resonator, and the energy stored in the electric field at the voltage maximum are both

$$W_s = (1/2)V^2C \tag{a14}$$

$$= (1/2)V^2M/\omega_0 \tag{a15}$$

Here V is the peak gap voltage. Expressions (a14) and (a15) are valuable in making resonator calculations from exact or approximate field distributions. They define C, L and M in terms of electric and magnetic field. The energy dissipated per cycle is

$$W_L = \pi V^2 C/\omega_0 \tag{a16}$$

Hence, we might have written Q as

$$Q = 2\pi W_s/W_L \tag{a17}$$

This is one popular definition of Q.

In (a8)–(a17) we usually assume that there is no appreciable energy stored in the load or the field of the coupling loop, so that M is considered as unaffected by load. The effect of "high Q" loads with considerable energy storage is considered in a somewhat different manner in Sec. IXB.

It must be emphasized that the expressions given above are valid for high Q circuits only (a Q of 50 is high in this sense). Expression (a17) is often used as a general definition of Q, but it is not complete without an additional definition of the meaning of resonance in a low Q circuit with many modes. Schelkunoff uses another definition of Q. Unforced oscillations in a damped circuit can be represented as a combination of several terms

$$V_1 e^{p_1 t} + V_2 e^{p_2 t} + \cdots \tag{a18}$$

$$p_1 = \alpha_1 + j\omega_1 \tag{a19}$$

Schelkunoff takes the Q of the nth mode as

$$Q_n = \omega_n/\alpha_n \tag{a20}$$

This is at least a consistent and complete definition. The reader can easily see that it accords with the definitions given for high Q's in connection with the circuit of Fig. 116.

Sometimes there may be a complicated circuit between the gap and a coaxial line or wave guide. In this case, the circuits intervening between the gap and the line can be regarded as a 4 terminal transducer (Fig. 117). The constants of this transducer will vary with frequency. No further consideration of this generalized treatment will be given, as it is well covered in books on network theory. A particular representation of the transducer will be pointed out, however. If the impedance in the line is referred to a special point, one parameter can be eliminated, giving the equivalent circuit shown in Fig. 118. If the gap is short-circuited, the impedance is zero and the impedance at the special point to be chosen on the line is R; the special point may be chosen as the potential minimum with the gap shorted.

N, the voltage ratio of a perfect transformer, is usually complex. The impedance ratio NN^* is real. If we are not interested in the relation of output phase to gap phase, we may disregard the phase angle of N, and deal only with the impedance ratio, the absolute value of N squared, which we will call N^2. Thus, using this equivalent circuit and disregarding the phase of N we can for our purposes reduce the number of independent parameters from the usual 6 for a passive 4 terminal network to 4, $Y(= G + jB)$, N^2 and R. This reduction can greatly simplify the algebra and arithmetic of microwave problems. The circuit of Fig. 118 has an additional advantage; if we choose our impedance reference point on the output line or guide to be the suitable point nearest to the actual output

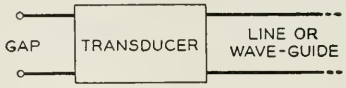


Fig. 117.—A 4-terminal transducer is the most general connection between the oscilla tor gap and a line or wave guide.

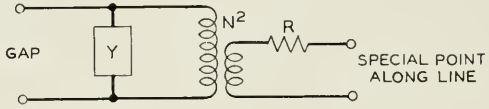


Fig. 118.—One circuit which will represent all the properties of a general 4-terminal transducer. This circuit consists of an admittance Y shunting the gap, a perfect transformer of complex ratio N and a series resistance R.

loop or iris, the circuit represents fairly accurately the frequency dependence of the output impedance if we merely take Y as

$$Y = G_R + j\omega(C - 1/\omega L) \tag{a21}$$

Near resonance we may use the simpler form.

$$Y = G_R + j2M\Delta\omega/\omega_0 \tag{a22}$$

Something has already been said in a general way about the evaluation of L and C in terms of the electric and magnetic field distribution in a resonator. It is completely outside of the scope of this paper to consider this subject at any length; the reader is referred to various books.[19, 20, 21] The discontinuity calculations of Whinnery and Jamieson[22] are also of great value in

[19] Electromagnetic Waves, S. A. Schelkunoff, Van Nostrand, 1943.
[20] Fields & Waves in Modern Radio, Ramo & Whinnery, Wiley, 1944.
[21] Microwave Transmission Data, Sperry Gyroscope Company, 1944.
[22] J. R. Whinnery and H. W. Jamieson, Equivalent Circuits for Discontinuities in Transmission Lines, *Proc. I.R.E.*, 32, 2, pp. 99–114.

making resonator calculations. These can be profitably combined with disk transmission line formulae.[19, 20]

The writers would like to point out that in the present state of the art the testing of resonator calculations by models is important. Models need not be made of the size finally desired. If all dimensions are made N times as large, the wavelength will be N times as great. The characteristic admittance M will be unchanged. If the material is the same, and the surface is smooth and homogeneous, Q and $1/G_R$, the shunt resonant resistance, will be $\sqrt{N}$ times as great.

It is perhaps a needless caution to say that the accuracy of a method of resonator calculation cannot be judged by its mathematical complexity or the difficulty of using it. Methods of calculation which are simple and may seem to make unduly broad approximations are sometimes better founded than appears on the surface, and complicated methods, exact if carried far enough, may be so unsuited to the problem as to give very bad answers if used in obtaining approximate values.

APPENDIX II

Modulation Coefficient

In this appendix the effects of space charge are neglected.

The modulation coefficient β is defined as the peak energy in electron volts an electron can gain in passing through the field of a gap divided by the peak r-f voltage across the gap. If an electron were transported across the gap very quickly when the r-f voltage was at a maximum the energy in electron volts gained by the electron would be equal to the peak r-f voltage. Thus, the modulation coefficient can also be defined as the ratio of the peak energy actually gained to the energy which would be gained in a very quick transit at the time of maximum voltage.

In this appendix modulation coefficient will be considered only for r-f voltages small compared with the d-c accelerating voltage.

If an electron gains an energy β times the *r-f* voltage V across the gap, the work done on it is βeV electron volts. By the conservation of energy, an induced current must flow between the electrodes of the gap, transferring a charge-βe against the voltage V and hence taking an amount of energy βeV from the circuit. Pursuing this argument we see that the modulation coefficient β times the electron convection current in the beam, q, gives the current induced in the gap by electron flow. In a circuit sense, there is fed into the gap, as from an infinite impedance source, an induced current βq.

We will assume that the gap involves a region in which the field along the electron path rises from zero and falls to zero again. This region is assumed

to be small compared with a wavelength and to have little a-c magnetic field in it, so that we can pretty accurately represent the field in this restricted region as the gradient of a potential. Along the path the potential is taken as the real part of

$$V(x)e^{j\omega t} \tag{b1}$$

For small gap voltages, to first order, the time that an electron reaches a given position may be taken as unaffected by the signal, so that

$$t = x/u_0 + t_0 \tag{b2}$$

Then the gradient along the path is

$$\frac{\partial V}{\partial x} = \text{Real } V'(x)e^{j(\omega x/u_0+\omega t_0)}. \tag{b3}$$

The change in momentum in passing through the field may be obtained by integrating the force on an electron times the time through the field. Let points a and b be in the field-free region to the left and right of the gap. Then we have

$$\Delta(\mathring{x}) = \text{Real } \frac{\eta}{u_0}\int_a^b V'(x)e^{j(\omega x/u_0+\omega t_0)}\,dx \tag{b4}$$

$$\Delta(\mathring{x}) = \text{Real } \frac{\eta e^{j\omega t_0}}{u_0}\int_a^b V'(x)e^{j\gamma x}\,dx \tag{b5}$$

$$\gamma = \omega/u_0\,. \tag{b6}$$

The integral will be a complex quantity. The exponential factor involving the starting time t_0 will rotate this. $\Delta(\mathring{x})$ will have a maximum value when the rotation causes the vector to lie along the real axis, and this maximum value is thus the absolute value of the integral. Hence

$$\Delta(\mathring{x})_{\max} = \frac{\eta}{u_0}\left|\int_a^b V'(x)e^{j\gamma x}\,dx\right|. \tag{b7}$$

For a-c voltages small compared with the voltage specifying the speed u_0, the energy change is proportional to the momentum change. For an electron transported instantly from one side of the gap to the other, the momentum change can be obtained by setting $\gamma = 0$.

$$\begin{aligned}\Delta(\mathring{x})_0 &= \frac{\eta}{u_0}\int_a^b V'(x)\,dx \\ &= \frac{\eta}{u_0}V.\end{aligned} \tag{b8}$$

Here V is the voltage between a and b. Hence, the modulation coefficient β is given for small signals by

$$\beta = (1/V)\left|\int_a^b V'(x)e^{j\gamma x}\,dx\right| \tag{b9}$$

$$V = V(b) - V(a) \tag{b10}$$

$$\gamma = \omega/u_0 \tag{b11}$$

$$u_0 = \sqrt{2\eta V_0} \tag{b12}$$

Thus u_0 is the electron velocity.

It is sometimes convenient to integrate by parts, giving the mathematical expression for β a different form

$$\beta = (1/V)\left|\frac{V'(x)e^{j\gamma x}}{j\gamma}\Bigg|_a^b - \frac{1}{j\gamma}\int_a^b V''(x)e^{j\gamma x}\,dx\right|$$

as $V'(x)$ is zero at a and b

$$\beta = (1/V)\left|\frac{1}{\gamma}\int_a^b V''(x)e^{j\gamma x}\,dx\right|. \tag{b13}$$

An interesting and important case is that of a uniform field between grids. Let the first grid be at $x = 0$ and the second at $x = d$. There is an abrupt transition to a gradient V/d at $x = 0$, and another abrupt transition to zero gradient at $x = d$. Thus, the integral (b13) is reduced to these two contributions, and we obtain

$$\beta = (1/V)\left|\frac{V}{\gamma d}\left(1 - e^{j\gamma d}\right)\right|. \tag{b14}$$

This is easily seen to be

$$\beta = \sin(\gamma\, d/2)/(\gamma\, d/2) \tag{b14}$$

This function, the modulation coefficient for fine parallel grids, is plotted in Fig. 119.

Sometimes apertures, as, circular apertures, or long narrow slits are used without grids. There are important relations between the modulation coefficient for a path on the axis and one parallel to the axis for such systems.[23]

In a two-dimensional gap system with axial symmetry, if the modulation coefficient for a path along the axis is β_0, the modulation coefficient for a path y away from the axis is

$$\beta_y = \beta_0 \cosh \gamma y \tag{b15}$$

[23] These relations first came to the attention of the writers through unpublished work of D. P. R. Petrie, C. Strachey and P. J. Wallis of Standard Telephones and Cables.

In an axially symmetrical electrode system, if the modulation coefficient on the axis is β_0, the modulation coefficient at a radius r is

$$\beta_r = \beta_0 I_0 (\gamma r) \tag{b16}$$

Here I_0 is a modified Bessel function.

It is easy to see why (b15) and (b16) must be so. The field in the gap can be resolved by means of a Fourier integral into components which vary

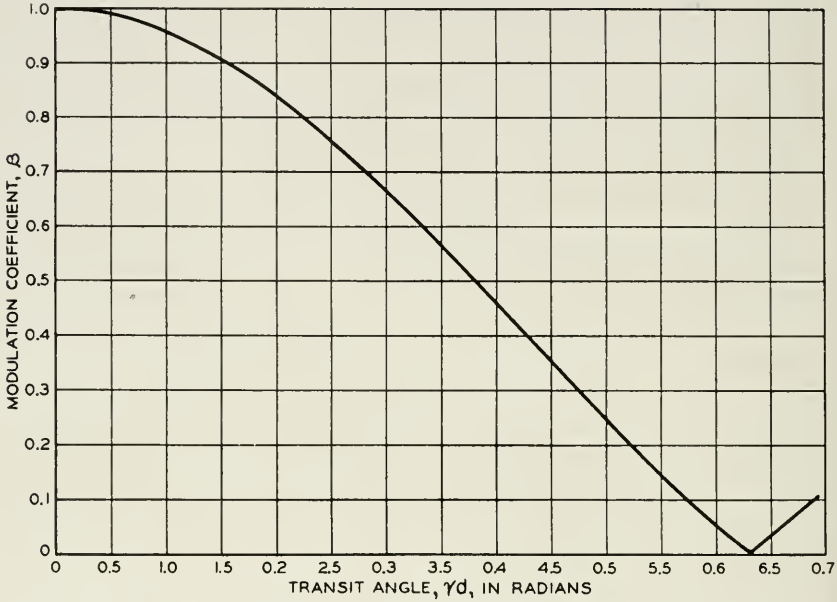


Fig. 119.—Modulation coefficient for fine parallel grids vs transit angle across the gap in radians.

$$\beta = |\sin(\gamma d/2)/(\gamma d/2)|, \qquad \gamma = \omega/v_0 = 3170/\lambda\sqrt{V_0}.$$

sinusoidally along the axis and as the hyperbolic cosine (in the two-dimensional case) of the same argument normal to the axis or as the modified Bessel function (in the axially symmetrical case) of the same argument radially. When the integration of (b9) is carried out, only that portion of the Fourier integral representation for which the argument is γx contributes to the result, and as that part contains as a factor cosh γx or $I_0(\gamma x)$ (b15) and (b16) are established.

The simple theory of velocity modulation presented in Appendix III makes no provision for variation of modulation coefficient across the beam. If we confine ourselves to very small signals, we find that the factor which appears is β^2. We may distinguish two cases: If the distance from the axis

of symmetry were the same for both transits of the electron, we would do well to average β^2. If the electrons got thoroughly mixed up in position between their two transits, we would do well to average β and then square the average. We will present both average and r.m.s. values of β. The averaged value of β will be denoted as β_a, the r.m.s. value as β_s; the value on the axis will be called β_0.

From (b15) and (b16) we obtain by simple averaging for the two dimen sional case,

$$\beta_a = \beta_0 \frac{\sinh \gamma y}{\gamma y} \tag{b17}$$

$$\beta_s = \beta_0 \left[\tfrac{1}{2} \left(\frac{\sinh 2\gamma y}{2\gamma y} + 1 \right) \right]^{\frac{1}{2}} \tag{b18}$$

and for an axially symmetrical case

$$\beta_a = \beta_0 2I_1(\gamma r)/(\gamma r) \tag{b19}$$

$$\beta_s = \beta_0 [I_0^2(\gamma r) - I_1^2(\gamma r)]^{1/2} \tag{b20}$$

It is convenient to rewrite these in a slightly different form, using (b15) and (b16).

$$\beta_a = \beta_y (\tanh \gamma y/\gamma y) \tag{b21}$$

$$\beta_s = \beta_y \left[\frac{\sinh 2\gamma y + 2\gamma y}{2\gamma y(\cosh 2\gamma y + 1)} \right]^{\frac{1}{2}} \tag{b22}$$

$$\beta_a = \beta_r 2I_1(\gamma r)/(\gamma r) I_0(\gamma r) \tag{b23}$$

$$\beta_s = \beta_r [1 - I_1^2(\gamma r)/I_0^2(\gamma r)]^{1/2} \tag{b24}$$

Now consider two similar cases: two pairs of parallel semi-infinite plates with a very narrow gap between them, and two semi-infinite tubes of the same diameter, on the same axis and with a very narrow gap between them (see Fig. 120). For electrons traveling very near the conducting surface, V' is zero save over a very short range at the gap, and the modulation coefficient is unity. Thus, by putting $\beta_y = \beta_r = 1$, we can use expressions (b15)–(b24) directly to evaluate β_a, β_s and β_0 for the configurations described. These quantities are shown in Figs. 121 and 122.

Suppose, now, that the gap between the plates or tubes is not very small. In this case, we need to know the variation of potential with distance across the space d long which separates the edges of the gap in order to get the modulation coefficient at the very edge of the gap, β_y or β_r.

If the tubing or plates surrounding the gap are thick, we might reasonably

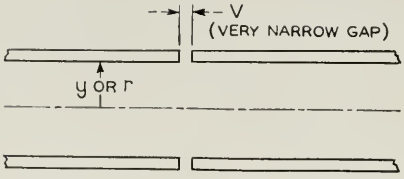


Fig. 120.—A gap consisting of pairs of semi-infinite planes or semi-infinite tubes.

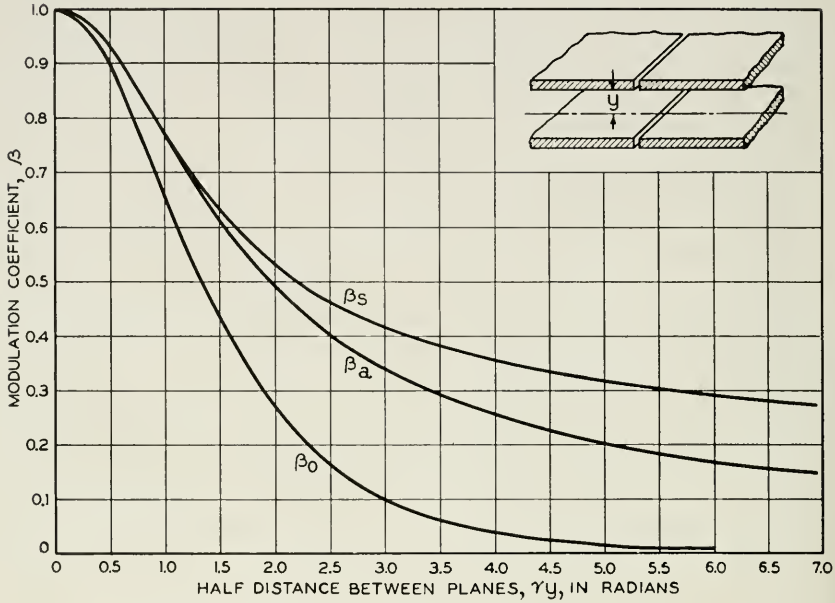


Fig. 121.—Modulation coefficient for two semi-infinite pairs of parallel planes with the edges very very close together, plotted vs the half distance between planes in radians. β_0 is the modulation coefficient for electrons travelling along the axis. β_a is the average modulation coefficient and β_s is the r.m.s. modulation coefficient. The separation of the planes is $2y$, $\beta_0 = 1/\cosh \gamma y$,

$$\beta_a = \tanh \gamma y/\gamma y, \qquad \beta_s = \left[\frac{\sinh 2\gamma y + 2\gamma y}{2\gamma y(\cosh 2\gamma y + 1)}\right]^{\frac{1}{2}}$$

assume a linear variation of potential with distance in the space between them. In this case, (b9) gives

$$\beta_y \text{ or } \beta_r = F_1(\gamma d) = \sin(\gamma d/2)/(\gamma d/2) \tag{b25}$$

This is the same function shown in Fig. 119.

If the tube wall or plates are very thin, one may, following Petrie, Strachey and Wallis[23] assume a potential variation between the edges of the gap of

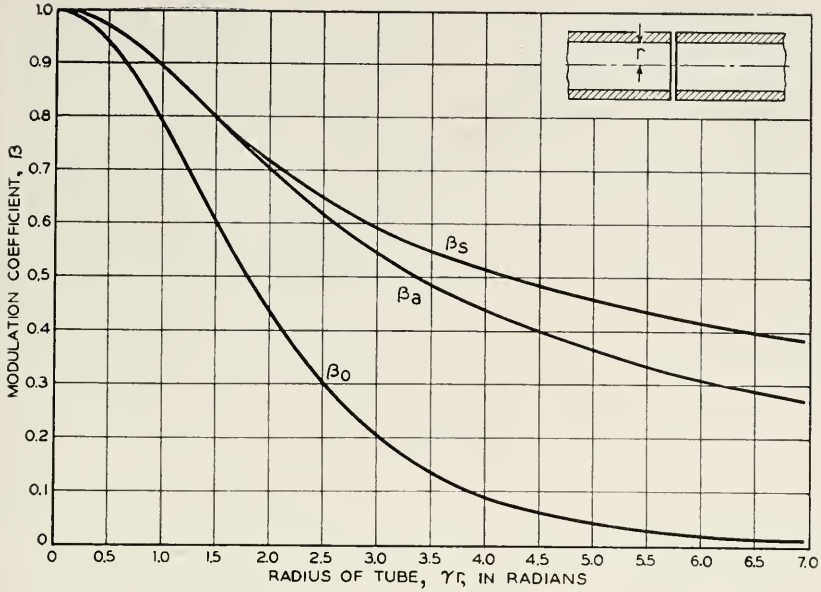


Fig. 122.—Modulation coefficient for two semi-infinite tubes separated by a very small distance, plotted vs the radius of the tube in radians. β_0 is the modulation coefficient on the axis, β_a is the average modulation coefficient and β_s is the root mean square modulation coefficient. r is the radius of the cylinders.

$$\beta_0 = 1/I_0(\gamma r), \qquad \beta_a = 2I_1(\gamma r)/\gamma r I_0(\gamma r),$$
$$\beta_s = [1 - I_1^2(\gamma r)/I_0^2(\gamma r)]^{\frac{1}{2}}$$

the form

$$V = \frac{1}{\pi} \sin^{-1} \frac{2x}{d} \tag{b26}$$

In this case, (b9) gives

$$\beta_y \text{ or } \beta_r = F_2(\gamma d) = J_0(\gamma d/2) \tag{b27}$$

Both $F_1(\gamma d)$ and $F_2(\gamma d)$ are plotted vs. γd in Fig. 123.

Figures 121, 122 and 123 cover fairly completely the case of slits and holes. The same methods may be used to advantage in making an approximate calculation taking into account the effect of grid pitch and wire size on modulation coefficient.

Assume we have a pair of lined up grids, as shown in Fig. 124. Approximately, the potential near the left one is given as

$$V = V_1' x/2 + (aV_1'/4\pi) \ln \left[2 \left(\cosh \frac{2\pi x}{a} - \cos \frac{2\pi y}{a} \right) \right]. \tag{b28}$$

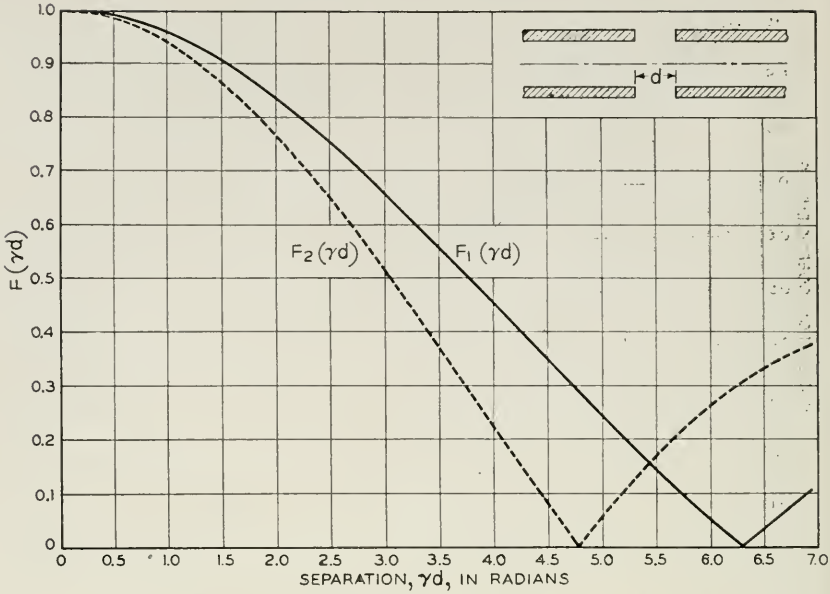


Fig. 123.—If the planes or tubes considered in Figs. 121 and 122 are separated by appreciable distance, the modulation coefficients given in those figures must be multiplied by a factor $F(\gamma d)$. In this figure, the multiplying factor is evaluated and plotted vs the separation in radians for two assumptions—that the gap has very blunt edges ($F_1(\gamma d)$) and that the gap has very sharp edges ($F_2(\gamma d)$).

This is zero far to the left and $V_1'x$ far to the right. This expression is useful only when the wire radius r is quite small compared with the separation a. Midway between wires,

$$V = V_1' x/2 + (a V_1'/2\pi) \ln\left[2 \cosh \frac{\pi x}{a}\right] \tag{b29}$$

$$V'' = (V_1' \pi/2a) \operatorname{sech}^2 \frac{\pi x}{a} \tag{b30}$$

If we use (b30) for each grid, the values of V'' for the two grids will overlap somewhat. However, let us neglect this overlap, apply (b30) at each grid, and using (b13), integrate for each grid from $-\infty$ to $+\infty$, giving

$$\beta_0 = (1/V)(V_1' \pi/\gamma a) \mid 1 - e^{j\gamma d} \mid \left| \int_{-\infty}^{\infty} \operatorname{sech}^2 \frac{\pi x}{a} e^{j\gamma x} \, dx \right| = f\{\sin(\gamma d/2)/(\gamma d/2)\} G_0(\gamma a) \tag{b31}$$

where

$$f = V_1'/(V/d) \tag{b32}$$

$$G_0(\gamma a) = \tfrac{1}{2} \left| \int_{-\infty}^{\infty} e^{(j\gamma a/\pi)u} \operatorname{sech}^2 u \, du \right| = (\gamma a/2)/\sinh(\gamma a/2) \tag{b33}$$

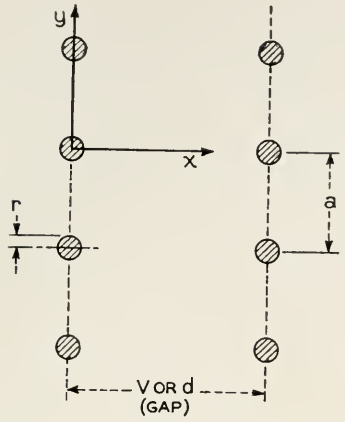


Fig. 124.—A gap consisting of lined up grids.

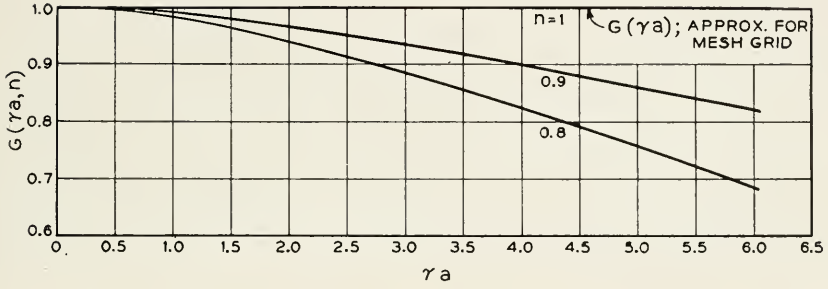


Fig. 125.—A factor used in obtaining the modulation coefficient of lined up grids vs the wire spacing in radians. n is the fraction open. The curve for $n = 1$ also applies approximately for a mesh grid.

Suppose we average over the open space of the grid. If the grid is a fraction n open, from (b17) we see that the average over the open space will be obtained by substituting for G_0 a quantity $G(\gamma a, n)$ given by

$$G(\gamma a,n) = \{\sinh (n\gamma a/2)/(n\gamma a/2)\}G_0(\gamma a) \qquad \text{(b34)}$$

In Fig. 125, $G(\gamma a,n)$ is plotted vs. γa for $n = 1, .9, .8$. This about covers the useful range of values. It should be positively noted that the average is over the *open* area of the grid and applies to current getting through.

It remains to evaluate the factor f. Suppose $x = 0$. At the surface of the grid wire, $y = r$, the radius of the wire,

$$\delta V = (V_1' a/4\pi) \ln\left[2\left(1 - \cos\frac{2\pi r}{a}\right)\right] = (V_1' a/2\pi) \ln\left[2 \sin\frac{\pi r}{a}\right]$$

As we have already assumed r is small, we may as well write

$$\delta V = (V_1' a/2\pi)2.3 \log_{10}\left(\frac{a}{2\pi r}\right) \tag{b35}$$

This emphasizes the sign of δV.

According to (b28), the grid plane appears from a distance to be at zero potential. Thus,

$$V_1' d - 2\delta V = V \tag{b36}$$

and from (b32)

$$f = \{1 + (a/\pi\ d)\ 2.3 \log_{10}(a/2\pi r)\}^{-1} \tag{b37}$$

If we go back over our results, we have for lined-up singly-wound grids, from (b31), (b34) and (b37), the average modulation coefficient

$$\beta_a = f\{\sin(\gamma a/2)/(\gamma a/2)\}G(\gamma a,n) \tag{b38}$$

The quantity $\sin(\gamma a/2)$ can be obtained from Fig. 119, $G(\gamma a,n)$ is plotted in Fig. 125, and f can be calculated from (b37) above.

It must be emphasized again that these expressions are good only for very fine wires ($r \ll a$), and get worse the closer the spacing compared with the wire separation. It is also important to note that $G(\gamma a,n)$ indicates little reduction of β_a even for quite wide wire separation. Now γd will be less than 2π, as $\beta_a = 0$ at $\gamma d = 2\pi$. As a approaches d in magnitude, the assumptions underlying the analysis, in which the integration around each grid was carried from $-\infty$ to ∞, become invalid and the analysis is not to be trusted.

It is very important to bear one point in mind. If we design a resonator assuming parallel conducting planes a distance L apart at the gap, and then desire to replace these planes with grids without altering the resonant frequency, we should space the grids not L apart but

$$d = fL \tag{b39}$$

apart to get the same capacitance and hence the same resonant frequency.

Mesh grids are sometimes used. To get a rough idea of what is expected, we may assume the potential about a grid to be

$$\begin{aligned} V = V_1' x/2 &+ (aV_1'/8\pi) \ln\left[2\left(\cosh\frac{2\pi x}{a} - \cos\frac{2\pi y}{a}\right)\right] \\ &+ (aV_1'/8\pi) \ln\left[2\left(\cosh\frac{2\pi y}{a} - \cos\frac{2\pi z}{a}\right)\right] \end{aligned} \tag{b40}$$

Here the grid is assumed to lie in the y, z, plane. This gives a mesh of wires about squares a on a side, the wires bulging at the intersections. We take r to be the wire radius midway between intersections.

We see that β_0 will be the same in this case as in the case of a parallel wire grid. Thus the added wires, which intercept electrons, haven't helped us as far as this part of the expression goes.

As a further approximation, an averaging will be carried out as if the apertures had axial symmetry. Averaging will be carried out to a radius giving a circle of area a^2. The steps will not be indicated.

Further a factor analogous to f will be worked out. Again, the steps will not be indicated. The results are

$$\beta_a = g\{\sin(\gamma d/2)/(\gamma d/2)\}G_1(\gamma a) \tag{b41}$$

$$G_1(\gamma a) = 2I_1(\gamma a/\sqrt{\pi})/(\gamma a/\sqrt{\pi})G_0(\gamma a) \tag{b42}$$

$$g = 1 + (.365\ a/d)\ (\log_{10}(a/\pi r) - .69) \tag{b43}$$

The quantity $G_1(\gamma,a)$ is plotted in Fig. 125 for comparison with the parallel wire case. It should be emphasized that these expressions assume $r \ll a$, and that $G_1(\gamma a)$ is really only an estimate based on a doubtful approximation. The indications are, however, that the only beneficial affect of going from a parallel wire grid to a mesh with the same wire spacing lies in a small decrease in δV (a small increase in the mu of the grid), while by doubling the number of wires in the parallel wire grid, a can be halved, both raising mu and increasing $G(\gamma a,n)$.

APPENDIX III

Approximate Treatment of Bunching

We assume[24] that the conditions are as shown in Fig. 126 where the electron energy on first entering the gap is specified by the potential V_0. Across the gap there exists a radio frequency voltage, $V \sin \omega t$. The ratio of the energy gained by the electron in crossing the gap to the energy which it would gain if the transit time across the gap were zero is called the modulation coefficient and is denoted by a factor, β. We assume that the modulation coefficient is the same for all electrons. We also neglect the effects of space charge throughout. After leaving the gap the electrons enter an electrostatic retarding field of strength $E_0 = \dfrac{V_R + V_0}{l}$

[24] This analysis follows the method given by Webster, J. App. Phys. 10, July 1939, pp. 501–508.

such that the stream flow is reversed and caused to retraverse the gap. The round trip transit time, τ_a, in the retarding field when a signal exists across the gap is then given by

$$\tau_a = \frac{2\sqrt{2\eta(V_0 + \beta V \sin \omega t_1)}}{\eta E_0} \tag{c1}$$

where $\eta = \frac{e}{m} \times 10^7 = 1.77 \times 10^{15}$ in practical units and t_1 is the time of first entry of the gap. The time of return to the gap will be

$$t_2 = t_1 + \tau_a \tag{c2}$$

More accurately t_1 and t_2 are measured from the central plane of the gap in which case a second term should be added to (c1) corresponding to motion at

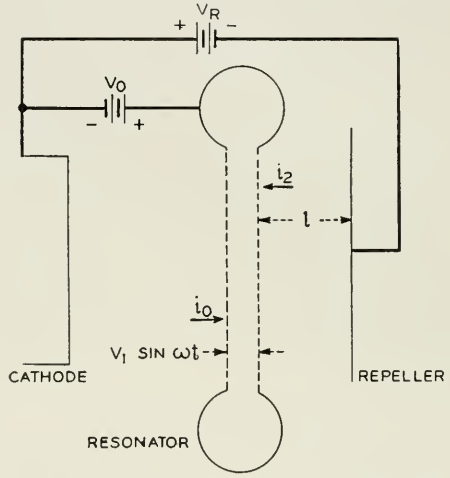


Fig. 126.—Diagram of a reflex oscillator showing quantities used in the treatment of bunching.

constant velocity. This term is, however, very small and will be neglected here. If i_2 is the current returning to the gap and I_0 the uniform current entering the gap on its first transit, then from conservation of charge one may write

$$I_0\, dt_1 = i_2\, dt_2 \tag{c3}$$

In what follows it will be first assumed that t_2 and t_1 are related by a single valued function. At the end of this appendix it will be shown that the analysis is also valid where the relating function is multiple valued.

We now make a Fourier series analysis of i_2 in order to determine the

harmonic distribution. Thus

$$i_2 = a_0 + a_1 \cos(\omega t_2 + \varphi) + a_2 \cos 2(\omega t_2 + \varphi) + \cdots + b_1 \sin(\omega t_1 + \varphi) + b_2 \sin 2(\omega t_2 + \varphi) + \cdots \tag{c4}$$

where

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} i_2 \cos n(\omega t_2 + \varphi)\, d\omega t_2$$
$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} i_2 \sin n(\omega t_2 + \varphi)\, d\omega t_2 \tag{c5}$$

Using (c1) to (c3) we change our variable to t_1 obtaining

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} I_0 \cos n\omega\left(t_1 + \frac{2\sqrt{2\eta(V_0 + \beta V \sin \omega t_1)}}{E_0} + \varphi\right) d\omega t_1 \tag{c6}$$

Let $\omega t_1 = \theta_1 \qquad \omega\tau_0 = \dfrac{2\omega\sqrt{2\eta V_0}}{E_0} = \theta \qquad X = \dfrac{\theta\beta V}{2V_0}$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} I_0 \cos n\left(\theta_1 + \varphi + \theta\left[1 + \frac{X}{\theta}\sin\theta_1 - \tfrac{1}{2}\frac{X^2}{\theta^2}\sin^2\theta_1 + \cdots\right]\right) d\theta_1 \tag{c7}$$

(c7) cannot be evaluated in closed form without further restriction. The first order theory may be obtained by assuming that $\frac{1}{2}\dfrac{nX^2}{\theta} \ll \dfrac{\pi}{2}$. It is not sufficient to assume that $\dfrac{X^2}{\theta} \ll X$. The latter assumes that the third and higher terms of the expansion are small compared to the second. Let the integrand be denoted as $i_0 \cos nx$. The quantity to be evaluated is the argument of a trigonometric function where the total angle is of less importance than the difference $nx - 2m\pi$ where m is the largest integer for which the difference is positive. The condition first expressed requires that the contribution of the third and higher terms to the difference phase shall be small. The restriction requires that

$$n\left(\frac{\beta V}{2V_0}\right)^2 \omega\tau_0 \ll \pi \tag{c8}$$

This is a more stringent requirement than

$$\frac{X}{\theta} = \frac{\beta V_1}{2V_0} \ll 1 \tag{c9}$$

(c9) requires only a small modulation depth while (c8) imposes a restriction on both the modulation depth and the drift time.

With the restriction (c8) imposed we obtain

$$a_n = \frac{1}{\pi}\int I_0 \cos n\left(\theta_1 + \varphi + \theta\left[1 + \frac{X}{\theta}\sin\theta_1\right]\right)d\theta_1 \tag{c10}$$

If we let $\varphi = -\theta$ all coefficients b_n will be zero and

$$a_n = 2(-1)^n I_0 J_n(Xn), \qquad a_0 = I_0 \tag{c11}$$

Thus the first order expansion for the current returning through the gap is

$$i_2 = I_0[(1 - 2J_1(X)\cos\omega(t_2 - \tau_0) + 2J_2(2X)\cos 2\omega(t_2 - \tau_0)\cdots] \tag{c12}$$

Our principal interest is in the fundamental component, which in complex notation is given by

$$(i_2)_f = -2I_0J_1(X)e^{j\omega(t_2-\tau_0)} \tag{c13}$$

It is shown in Appendix II that the circuit current induced in the gap will be given, if account is taken of the phase reversal of π resulting from the reversal of direction of the beam, by

$$I_2 = -\beta(i_2)_f$$

The gap voltage at the time of return will be $v = V\sin\omega t_2$ or in complex notation

$$v = Ve^{j(\omega t_2-(\pi/2))} \tag{c14}$$

Hence the electronic admittance to the fundamental will be

$$Y_e = \frac{I_2}{v} = \frac{2I_0\beta}{V}J_1\left(\frac{\omega\tau_0\beta V}{2V_0}\right)e^{-j(\omega\tau_0-(\pi/2))} \tag{c15}$$

In the foregoing it was assumed that t_2 was a single valued function of t_1. We may generalize by writing (c3) as

$$\sum I_0\,dt_1 = i_2\,dt_2 \tag{c16}$$

For sufficiently large signals there may be several intervals dt_1 which contribute charge to a given interval dt_2 and hence we write a summation for the left hand side of (c16). When the Fourier analysis is made and the change in variable from t_2 to t_1 is made the single integral breaks up into a sum of integrals. In Fig. 127 we plot time t_1 on a vertical scale with the sine wave indicating the instantaneous gap voltage. Displaced to the right on a vertical scale we plot time t_2. The solid lines connect corresponding times in the absence of signal for increments of time dt_1 and dt_2. When sufficiently

large signals are applied some of the electrons in the original interval dt_1 will gain or lose sufficient energy to be thrown outside the original corresponding interval dt_2 as for example as indicated by AB. If we consider a whole cycle of the gap voltage in time t_2 it is apparent that, under steady state conditions, for every electron which is thrown outside the corresponding cycle in t_2 another from a different cycle in t_1 is thrown in whose phase differs by a multiple of 2π as for example CD. In summing the effects of these charge increments the difference of 2π in starting phase produces no physical effect. This is of course also true mathematically in the Fourier analysis of a periodic function since in integrating over an interval 2π it is immaterial whether we integrate over a single interval or break it up into a

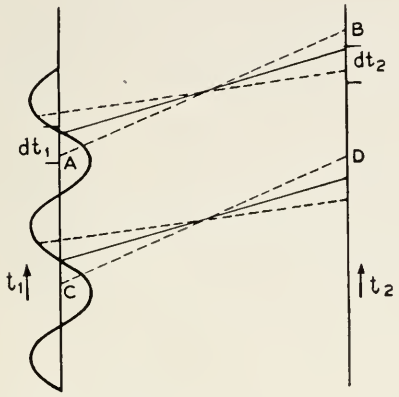


Fig. 127.—Diagram showing the relation between t_1, the time an electron crosses the gap for the first time, and t_2, the time the electron returns across the gap.

sum of integrals over intervals $-\pi$ to a, $2\pi n_1 + a$ to $2\pi n_1 + b$, $2\pi n_2 + b$ to $2\pi n_2 + c$, etc. where the subintervals sum up to 2π. Hence we conclude that the preceding analysis is also valid up to (c7) for signals sufficiently large so that t_2 and t_1 are related by a multiple valued function and is valid beyond that point provided that we do not violate (c8).

APPENDIX IV

Drift Angle as a Function of Frequency and Voltage

Let τ be the transit time in the drift space. Then the drift angle is

$$\theta = \omega\tau \tag{d1}$$

For changes in voltage (resonator or repeller), both τ and ω will change.

Thus

$$\begin{aligned}\Delta\theta/\theta &= \Delta\omega/\omega + \Delta\tau/\tau \\ &= \Delta\omega/\omega + ((\partial\tau/\partial V)/\tau)\Delta V\end{aligned} \tag{d2}$$

As shown in Appendix VI, the derivative of τ with respect to repeller voltage, $\partial\tau/\partial V_R$, is always negative, while the derivative of τ with respect to resonator voltage, $\partial\tau/\partial V_0$, may be either negative or positive. For a linear variation of potential in the drift region, $\partial\tau/\partial V_0$ is zero when $V_R = V_0$ and negative for smaller values of V_R.

APPENDIX V

Electronic Admittance—Non-Simple Theory

A closer treatment of the drift action in the repeller space follows, in which are considered the changes which occur as the voltage on the cavity becomes large.

The additional terms to be considered come from an evaluation in series of the higher-order terms of (c7), which were neglected in Appendix III. Only the fundamental component of current will be considered, although other terms could be included if desired. The integrals of interest may be rewritten from (c7), using the relation $\varphi = -\theta$, as follows:

$$a_1 = \frac{I_0}{\pi}\int_{-\pi}^{\pi}\cos\left(\theta_1 + X\sin\theta_1 - \frac{1}{2}\frac{X^2}{\theta}\sin^2\theta_1 + \frac{1}{2}\frac{X^3}{\theta^2}\sin^3\theta_1 + \cdots\right)d\theta_1 \tag{e1}$$

$$b_1 = \frac{I_0}{\pi}\int_{-\pi}^{\pi}\sin\left(\theta_1 + X\sin\theta_1 - \frac{1}{2}\frac{X^2}{\theta}\sin^2\theta_1 + \frac{1}{2}\frac{X^3}{\theta^2}\sin^3\theta_1 + \cdots\right)d\theta_1 \tag{e2}$$

We shall hereinafter neglect terms of higher order in $\frac{1}{\theta}$ than those explicitly shown here. With this neglect, we can expand the trigonometric functions, obtaining

$$\begin{aligned}a_1 = \frac{I_0}{\pi}\int_{-\pi}^{\pi}\cos(\theta_1 + X\sin\theta_1)\left[1 - \frac{1}{8}\frac{X^4}{\theta^2}\sin^4\theta_1 + \cdots\right]d\theta_1 \\ - \frac{I_0}{\pi}\int_{-\pi}^{\pi}\sin(\theta_1 + X\sin\theta_1) \\ \cdot\left[-\frac{1}{2}\frac{X^2}{\theta}\sin^2\theta_1 + \frac{1}{2}\frac{X^3}{\theta^2}\sin^3\theta_1 + \cdots\right]d\theta_1\end{aligned} \tag{e3}$$

$$b_1 = \frac{I_0}{\pi}\int_{-\pi}^{\pi}\sin(\theta_1 + X\sin\theta_1)\left[1 - \frac{1}{8}\frac{X^4}{\theta^2}\sin^4\theta_1 + \cdots\right]d\theta_1$$
$$+ \frac{I_0}{\pi}\int_{-\pi}^{\pi}\cos(\theta_1 + X\sin\theta_1) \tag{e4}$$
$$\cdot\left[-\frac{1}{2}\frac{X^2}{\theta}\sin^2\theta_1 + \frac{1}{2}\frac{X^3}{\theta^2}\sin^3\theta_1 + \cdots\right]d\theta_1$$

Now of these terms, not all give contributions; some integrate to zero since the integrand is an odd function of θ_1. Rewriting with those terms omitted,

$$a_1 = \frac{I_0}{\pi}\int_{-\pi}^{\pi}\cos(\theta_1 + X\sin\theta_1)\left[1 - \frac{1}{8}\frac{X^4}{\theta^2}\sin^4\theta_1 + \cdots\right]d\theta_1$$
$$- \frac{I_0}{\pi}\int_{-\pi}^{\pi}\sin(\theta_1 + X\sin\theta_1)\left[\frac{1}{2}\frac{X^3}{\theta^2}\sin^3\theta_1 + \cdots\right]d\theta_1 \tag{e5}$$

$$b_1 = \frac{I_0}{\pi}\int_{-\pi}^{\pi}\cos(\theta_1 + X\sin\theta_1)\left[-\frac{1}{2}\frac{X^2}{\theta}\sin^2\theta_1 + \cdots\right]d\theta_1 . \tag{e6}$$

Evaluation of these terms is formally simplified by the following relationships, each obtained by differentiation of the previous one:

$$-2J_1(X) = \frac{1}{\pi}\int_{-\pi}^{\pi}\cos(\theta_1 + X\sin\theta_1)\,d\theta_1 \tag{e7}$$

$$-2J_1'(X) = -\frac{1}{\pi}\int_{-\pi}^{\pi}\sin(\theta_1 + X\sin\theta_1)\sin\theta_1\,d\theta_1 \tag{e8}$$

$$-2J_1''(X) = -\frac{1}{\pi}\int_{-\pi}^{\pi}\cos(\theta_1 + X\sin\theta_1)\sin^2\theta_1\,d\theta_1 . \tag{e9}$$

Continuation of this process gives all the terms of interest in (e5) and (e6). Hence

$$a_1 = I_0\left(-2J_1 + \frac{1}{4}\frac{X^4}{\theta^2}J_1^{iv} + \frac{X^3}{\theta^2}J_1''' + \cdots\right) \tag{e10}$$

$$b_1 = I_0\frac{X^2}{\theta}(-J_1'' + \cdots). \tag{e11}$$

Therefore the expression for the fundamental component of the beam current may be written as follows, passing to complex notation:

$$(i_2)_f \sim a_1\cos(\omega t_2 - \theta) + b_1\sin(\omega t_2 - \theta)$$
$$(i_2)_f = (a_1 - jb_1)e^{j(\omega t_2 - \theta)} \tag{e12}$$
$$= e^{j\omega t_2}e^{-j\theta}I_0\left(-2J_1 + j\frac{X^2}{\theta}J_1'' + \frac{1}{\theta^2}(X^3J_1''' + \tfrac{1}{4}X^4J_1^{iv})\right) + \cdots .$$

Following usual conventions, the real part of the complex expression represents the physical situation, and the exponential time function will usually be omitted in what follows.

Similarly to Appendix III, the induced current in the gap is

$$I_2 = -\beta(i_2)_f$$

$$= e^{j\omega t}\beta e^{-j\theta} I_0\left(2J_1 - j\frac{X^2}{\theta}J_1'' - \frac{1}{\theta^2}(X^3 J_1''' + \tfrac{1}{4}X^4 J_1^{iv})\right). \quad \text{(e13)}$$

The gap voltage is still the same, viz.

$$v \backsim V\sin\omega t_2 \backsim Ve^{j(\omega t_2-(\pi/2))} = +\frac{2V_0}{\beta}\frac{X}{\theta}e^{j(\omega t_2-(\pi/2))}. \quad \text{(e14)}$$

Accordingly, the electronic admittance to the fundamental will be

$$Y_e = \frac{I_2}{v} = \frac{I_0}{V_0}\beta^2\frac{\theta}{X}e^{j((\pi/2)-\theta)}\left(J_1 - \frac{jX^2}{\theta 2}J_1'' - \frac{1}{\theta^2}\left(\frac{X^3}{2}J_1''' + \frac{X^4}{8}J_1^{iv}\right)\right) \quad \text{(e15)}$$

The argument of J_1 and its derivatives is understood to be X. The derivatives can be evaluated in terms of J_0 and J_1 by repeated use of the Bessel function recurrence relations (see Jahnke-Emde, Funktionentafeln, p. 144). The result is

$$Y_e = \frac{I_0}{V_0}\beta^2\frac{\theta}{X}e^{j((\pi/2)-\theta)}\left(J_1 - \frac{j}{\theta}\left[\left(1-\frac{X^2}{2}\right)J_1 - \frac{X}{2}J_0\right] - \frac{1}{\theta^2}\left[\tfrac{1}{8}(X^2+X^4)J_1 - \tfrac{1}{4}X^3 J_0\right] + \cdots\right). \quad \text{(e16)}$$

The real part of this will be of interest; it is

$$G_e = +\frac{I_0}{V_0}\beta^2\frac{\theta}{X}\sin\theta\left(J_1 + \frac{\cot\theta}{\theta}\left[\left(1-\frac{X^2}{2}\right)J_1 - \frac{X}{2}J_0\right] - \frac{1}{\theta^2}\left[\tfrac{1}{8}(X^2+X^4)J_1 - \tfrac{1}{4}X^3 J_0\right]\right). \quad \text{(e17)}$$

The power generated by the electron stream is given by

$$P = -\tfrac{1}{2}G_e V^2$$

$$= -I_0 V_0\frac{2X}{\theta}\left(\sin\theta\left[J_1 - \frac{1}{8\theta^2}((X^2+X^4)J_1 - 2X^3 J_0)\right] + \frac{\cos\theta}{\theta}\left[\left(1-\frac{X^2}{2}\right)J_1 - \frac{X}{2}J_0\right]\right) + \cdots. \quad \text{(e18)}$$

Of this power, only a part is usefully delivered to the load admittance G_L, the rest being dissipated because of the circuit loss conductance G_R. The power lost in the circuit is

$$P_R = \tfrac{1}{2}G_R V^2 = \frac{2V_0^2 G_R}{\beta^2}\frac{X^2}{\theta^2} \tag{e19}$$

Accordingly, the useful power delivered to the load is

$$P_L = P - P_R \tag{e20}$$

A quantity of interest is the maximum useful power which can be obtained from the reflex oscillator; this is given by

$$\frac{\partial P_L}{\partial X} = 0 = \frac{\partial P_L}{\partial \theta} \tag{e21}$$

These two conditions are expressed by the next two equations.

$$2XJ_0 + \frac{1}{\theta}\cot\theta\,(-X^3J_0 - X^2J_1) + \frac{1}{\theta^2}\left((\tfrac{7}{4}X^3 - \tfrac{1}{4}X^5)J_0 + \left(-\frac{X^2}{2} - \tfrac{3}{2}X^4\right)J_1\right) + \frac{4V_0}{\beta^2 I_0}G_R\frac{X}{\theta\sin\theta} = 0 \tag{e22}$$

$$XJ_0 + (-4 + X^2)J_1 + \theta\cot\theta\cdot 2J_1 - \frac{\cot\theta}{\theta}((-\tfrac{1}{2}X^3 - 2X)J_0 + (4 - \tfrac{7}{4}X^2 + \tfrac{1}{4}X^4)J_1) + \frac{1}{\theta^2}(-\tfrac{3}{2}X^3J_0 + \tfrac{3}{4}(X^2 + X^4)J_1) - \frac{4V_0}{\beta^2 I_0}G_R\frac{X}{\theta\sin\theta} = 0 \tag{e23}$$

One may note that even if one sets $G_R = 0$ and neglects terms in $\frac{1}{\theta^2}$ the second of these equations is a little different from the corresponding one of Appendix III, so that a slightly different phase angle is predicted for maximum generated power. The result of Appendix III was

$$\theta\cot\theta = 1$$

predicting a phase angle θ slightly less than $\theta_n = (n + \frac{3}{4})2\pi$. However, the zero order result here is

$$\theta\cot\theta = 2 - \frac{X^2}{2} = -.892 \tag{e24}$$

predicting a phase angle a trifle larger than θ_n, in the approximation of Appendix III.

The equations (e22) and (e23) may look as if drastic measures would now be needed, but a parametric solution is surprisingly easy; one need only solve (e22) for G_R, and substitute back into the power expression (e20). The results are

$$G_R = -\frac{I_0}{V_0}\frac{\beta^2}{4}\theta \sin\theta\left(2J_0 + \frac{\cot\theta}{\theta}[-X^2J_0 - XJ_1] + \frac{1}{\theta^2}\left[(\tfrac{7}{4}X^2 - \tfrac{1}{4}X^4)J_0 + \left(-\frac{X}{2} - \tfrac{3}{2}X^3\right)J_1\right] + \cdots\right) \tag{e25}$$

$$P_L = -I_0V_0\frac{X}{\theta}\sin\theta\left(2J_1 - XJ_0 + \frac{\cot\theta}{\theta}\left[\left(-X + \frac{X^3}{2}\right)J_0 + \left(2 - \frac{X^2}{2}\right)J_1\right] + \frac{1}{\theta^2}\left[\frac{X^4}{2}J_1 + (-\tfrac{3}{8}X^3 + \tfrac{1}{8}X^5)J_0\right]\right) \tag{e26}$$

$$G_L = -G_e - G_R = -\frac{I_0}{V_0}\beta^2\frac{\theta}{X}\sin\theta\left(J_1 - \tfrac{1}{2}XJ_0 + \frac{\cot\theta}{\theta}\cdot\left[\left(1 - \frac{X^2}{4}\right)J_1 + \left(-\frac{X}{2} + \frac{X^3}{4}\right)J_0\right] + \frac{1}{\theta^2}[\tfrac{1}{4}X^4J_1 + (-\tfrac{3}{16}X^3 + \tfrac{1}{16}X^5)J_0]\right) \tag{e27}$$

One further convolution is necessary, because the equations (e25) and (e26) are still subject to the optimum phase angle condition (e23). Since we are here carrying only terms as far as $\frac{1}{\theta^2}$, approximations are in order.

From (e24) we get the hint that $\theta \cot\theta$ is of the order of unity, so that terms in $\frac{\cot\theta}{\theta}$ are of the same order as those in $\frac{1}{\theta^2}$. Accordingly an approximate solution is obtainable by adding (e22) and (e23) and neglecting these small terms. The result is

$$\theta\cot\theta = 2 - \frac{X^2}{2} - \frac{3XJ_0}{2J_1} \equiv F(X) \tag{e28}$$

Therefore the optimum phase angle is given by

$$\theta = \theta_n - \frac{F(X)}{\theta_n}, \qquad \theta_n = (n + \tfrac{3}{4})2\pi \tag{e29}$$

$$\sin\theta = -1 + \frac{[F(X)]^2}{2\theta_n^2} \tag{e30}$$

$$\frac{\sin\theta}{\theta} = -\frac{1}{\theta_n}\left[1 + \frac{1}{\theta_n^2}(F - F^2/2)\right] \tag{e31}$$

$$\frac{\cot\theta}{\theta} = \frac{1}{\theta_n^2}F(X) \tag{e32}$$

From computation it turns out that in the range of interest, the quantity $F(X)$ does not differ from (-1) by more than 20%.

The desired approximate solution comes now from substituting the explicit phase optimum (e29) to (e32) back into (e25)–(e27). The results are:

$$G_R = \frac{I_0}{V_0}\beta^2\frac{\theta_n}{2}\left(J_0 + \frac{1}{\theta_n^2}S_1(X)\right) \tag{e33}$$

$$P_L = I_0 V_0 \frac{X^2}{\theta_n}\left(\frac{2J_1}{X} - J_0 + \frac{1}{\theta_n^2}S_2(X)\right) \tag{e34}$$

$$G_L = \frac{I_0}{V_0}\beta^2\frac{\theta_n}{2}\left(\frac{2J_1}{X} - J_0 + \frac{1}{\theta_n^2}S_3(X)\right) \tag{e35}$$

The S-functions are given by

$$S_1(X) = \left(-F - \frac{F^2}{2} - \frac{X^2F}{2} + \frac{7}{8}X^2 - \frac{1}{8}X^4\right)J_0 + \left(-\frac{X^2}{2}F - \frac{X^2}{4} - \frac{3}{4}X^4\right)\frac{J_1}{X} \tag{e36}$$

$$S_2(X) = \left(-F^2 - F\frac{X^2}{2} + 4F + \frac{1}{2}X^4\right)\frac{J_1}{X} + \left(\frac{F^2}{2} + \frac{X^2F}{2} - 2F - \frac{3}{8}X^2 + \frac{1}{8}X^4\right)J_0 \tag{e37}$$

$$S_3(X) = \left(\frac{F^2}{2} + \frac{FX^2}{2} - \frac{3}{8}X^2 + \frac{1}{8}X^4\right)J_0 + \left(-F^2 - F\frac{X^2}{2} + \frac{1}{2}X^4\right)\frac{J_1}{X} \tag{e38}$$

The equations (e33)–(e35) have the following meaning: they presuppose that the load G_L has been adjusted for maximum useful power in the presence of circuit loss G_R, and that the drift angle is also optimum. Then the useful power is given parametrically in terms of the circuit conductance by equations (e33) and (e34), while (e35) gives the required optimum load conductance, also in terms of the parameter X.

The results may be expressed as a chart of useful power, plotted against the value of resonator loss conductance. This is done in Fig. 128.

One may also be interested in the maximum power which could be gen-

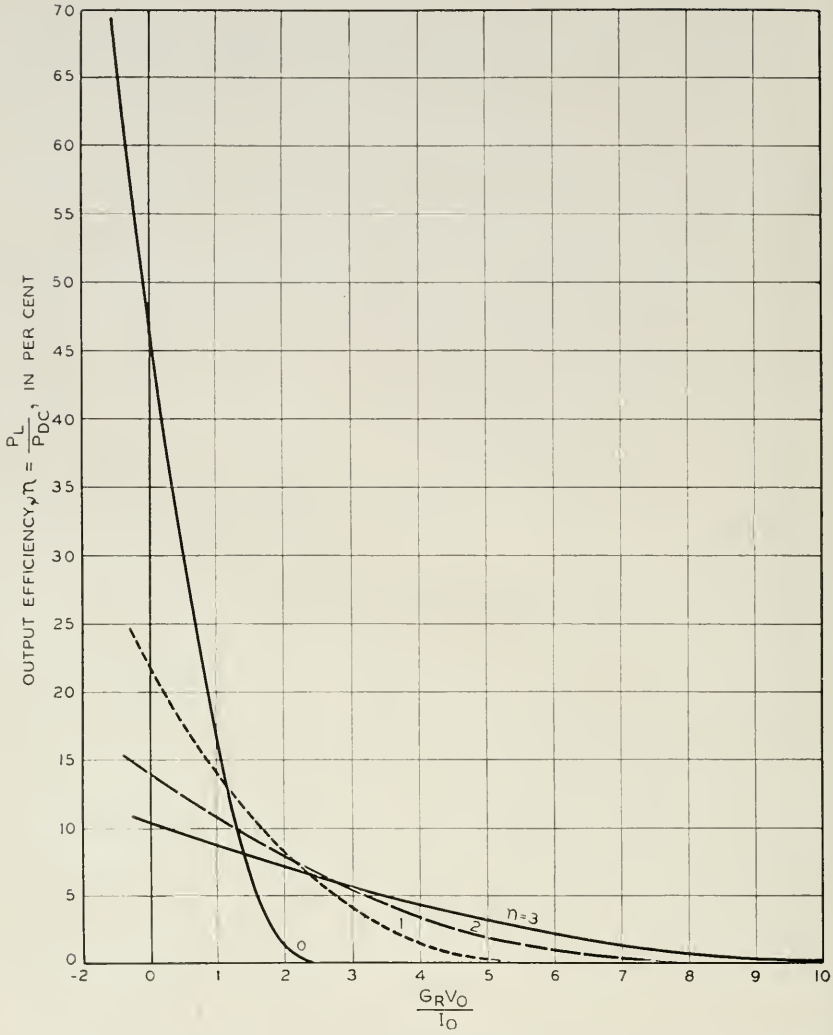


Fig. 128.—Plot of efficiency vs a parameter proportional to resonator loss for several repeller modes.

n	0	1	2	3
η	0.48	0.22	0.14	0.105

Fig. 129.—Maximum efficiency for several repeller modes.

erated if the resonator were perfect. This comes from setting $G_R = 0$, calculating the resulting value of useful power. The results are compared with the simpler theory in Fig. 129.

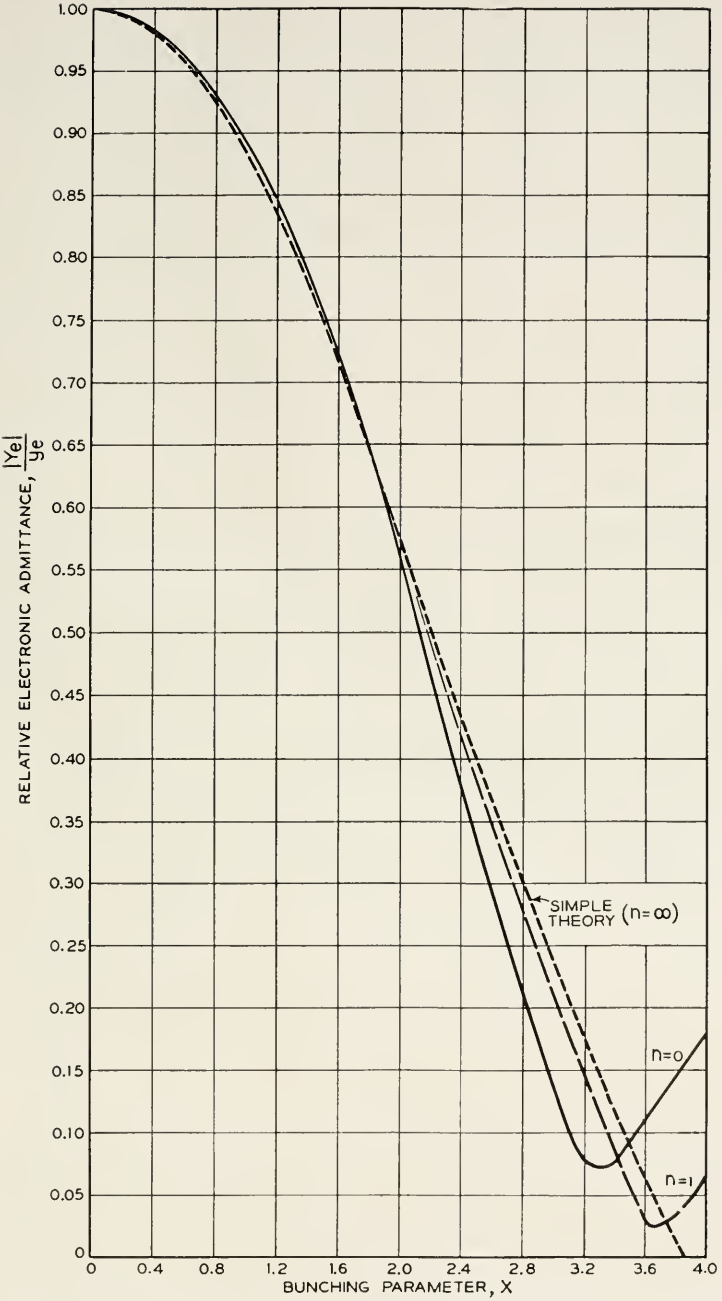


Fig. 130.—Relative electronic admittance vs bunching parameter for several repeller modes.

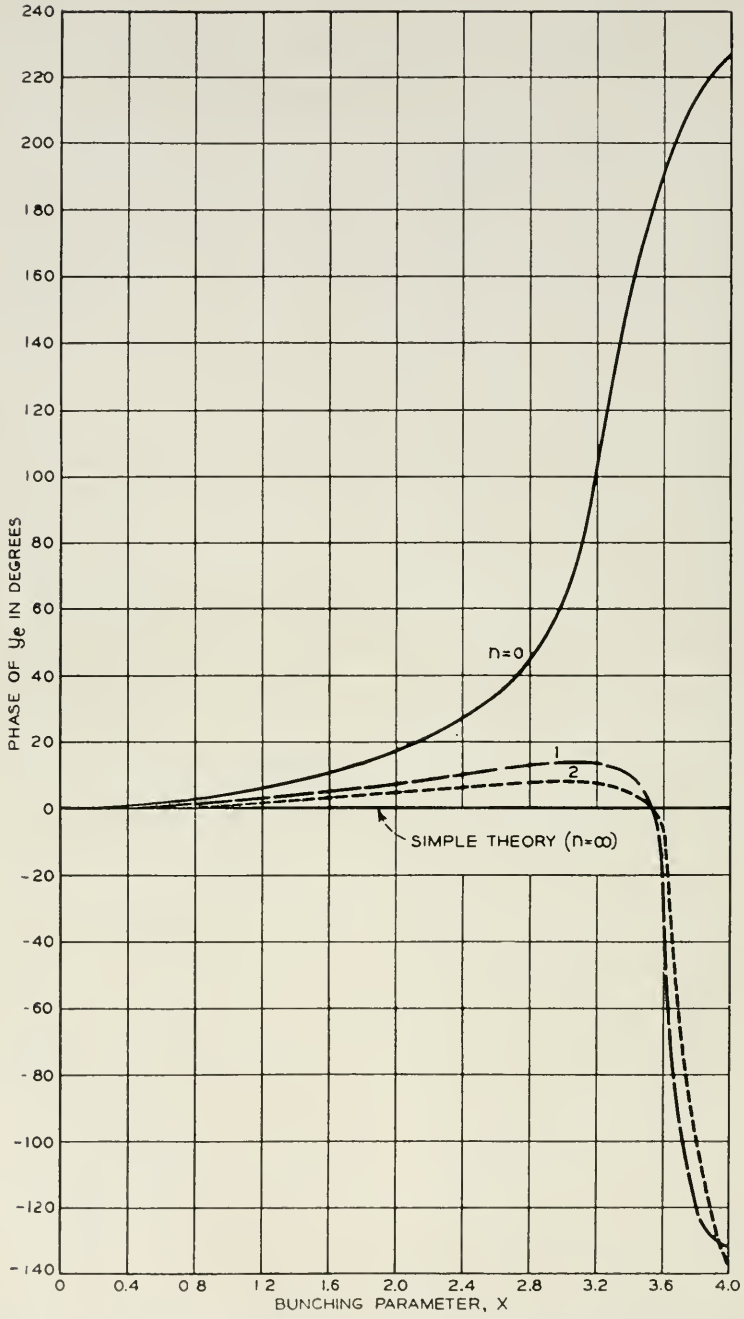


Fig. 131.—Phase of electronic admittance vs bunching parameter for several repeller modes.

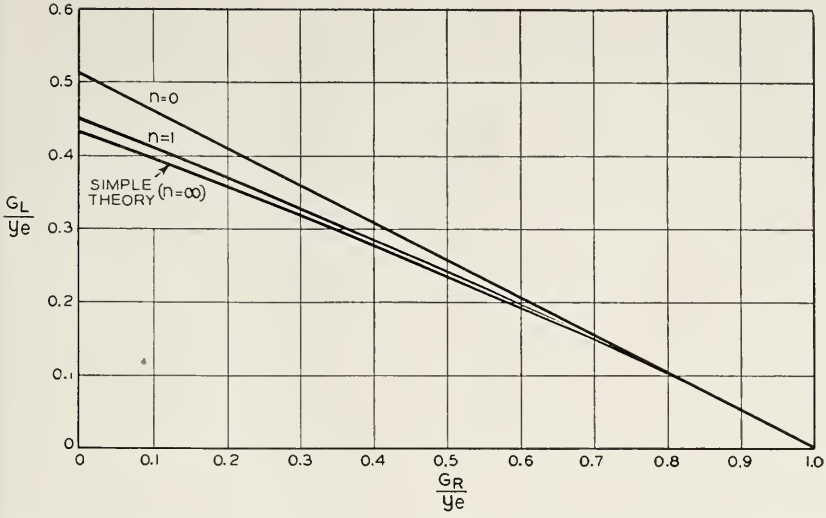


Fig. 132.—Optimum load conductance divided by small signal electronic admittance vs resonator loss conductance divided by small signal electronic admittance for several repeller modes.

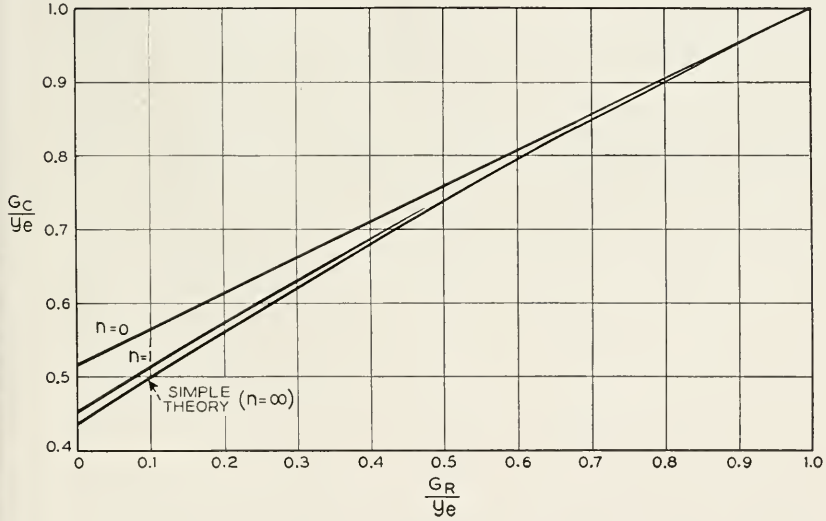


Fig. 133.—Optimum total circuit conductance divided by small signal electronic admittance vs resonator loss conductance divided by small signal electronic admittance for several repeller modes.

Further data which may be determined include the variation of the magnitude of the electronic admittance with gap voltage, in Fig. 130; also its phase, in Fig. 131. The optimum load conductance is plotted as a

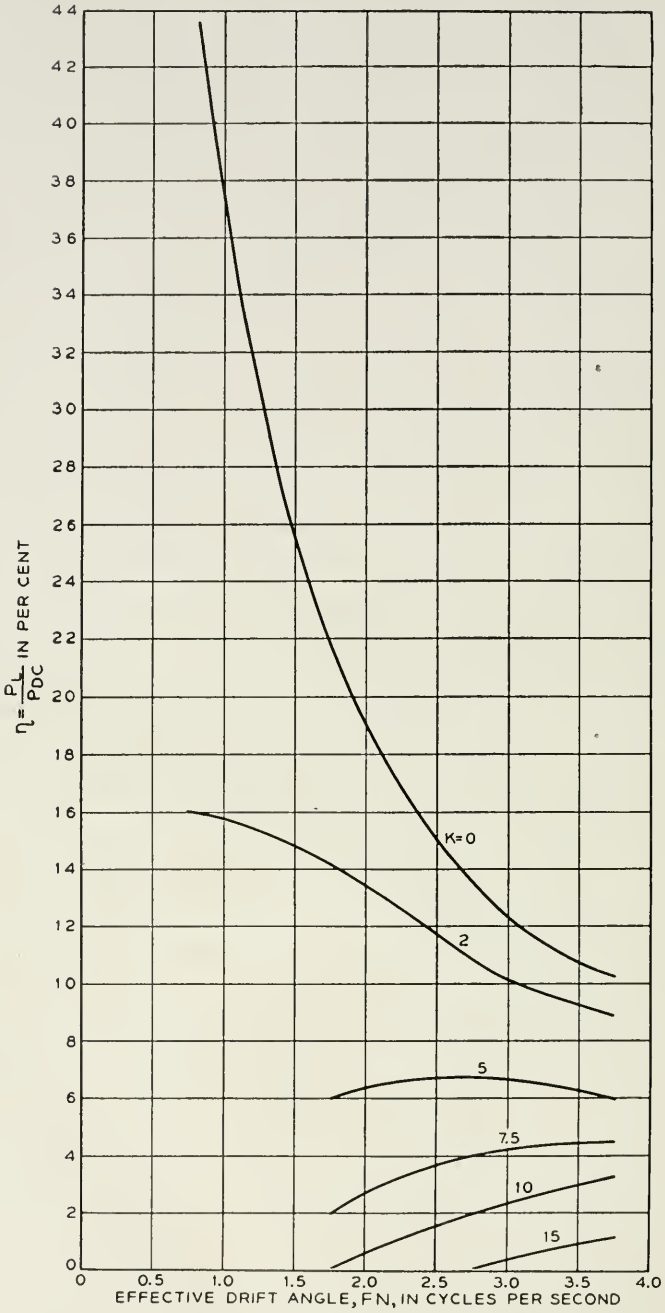


Fig. 134.—Efficiency vs effective drift angle for several degrees of resonator loss.

function of resonator loss in Fig. 132, and the total conductance, load plus loss, in Fig. 133. The next Fig. 134, plots efficiency versus mode number

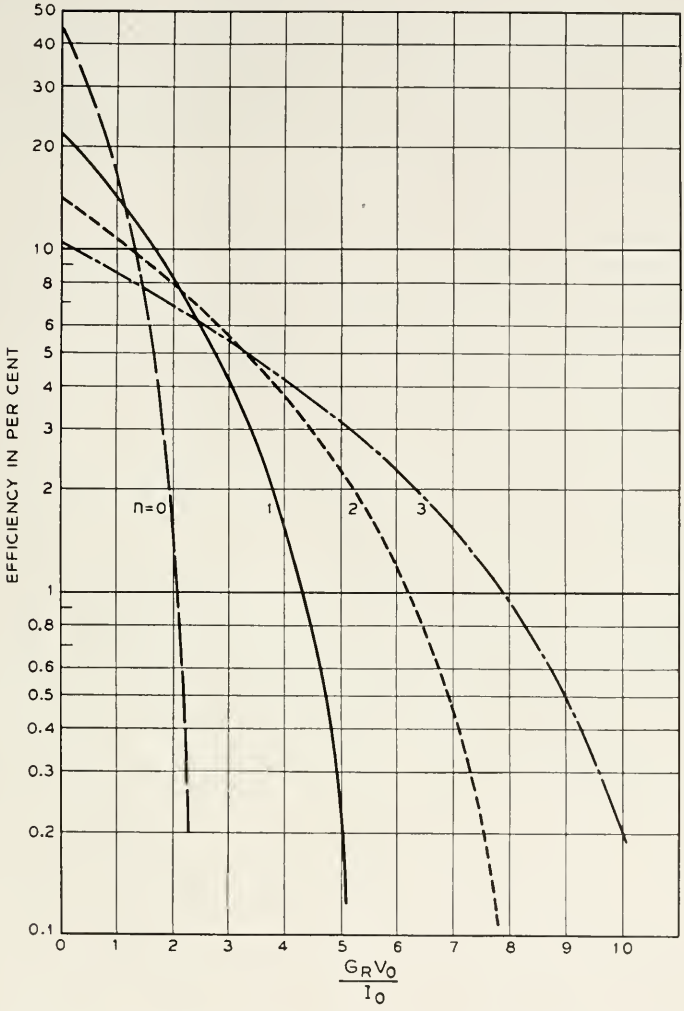


Fig. 135.—Efficiency vs a parameter proportional to resonator loss for several repeller modes.

with resonator loss as a parameter, while the next Fig. 135, plots efficiency versus resonator loss with mode number as a parameter.

Most of these graphs include for comparison the results of the simpler theory, and it can be seen that the deviations indicated by the second

order theory are ordinarily rather small even for the first two modes of operation, and are quite negligible for higher modes.

APPENDIX VI

General Potential Variation in the Drift Space

Suppose that the potential of the drift space is given by $V(x)$, where $x = 0$ at zero potential and $x = \ell$ at the gap. Then the transit time from the gap to zero potential and back again is

$$\tau_0 = (2/\sqrt{2\eta}) \int_0^\ell \frac{dx}{(V(X))^{\frac{1}{2}}} \tag{f1}$$

Imagine now that the entire drift space is raised by a very small amount ΔV. The zero potential point will now occur at

$$x = -\Delta V/V'(0) \tag{f2}$$

where

$$V'(x) = dV/dx \tag{f3}$$

Hence the new transit time will be

$$\tau_0 + \Delta\tau = (2/\sqrt{2\eta}) \int_{-\Delta V/V'(0)}^{\ell} \frac{dx}{[V(x) + \Delta V]^{\frac{1}{2}}} \tag{f4}$$

Now let

$$z = x + \Delta V/V'(0) \tag{f5}$$

Then, including first order terms only, if $V(x)$ can be expanded in a Taylor's series about 0,

$$\begin{aligned} \tau_0 + \Delta\tau &= (2/\sqrt{2\eta}) \int_0^{\ell+\Delta V/V'(0)} \frac{dz}{[V(z) - [V'(z)/V'(0)]\Delta V + \Delta V]^{\frac{1}{2}}} \\ &= (2/\sqrt{2\eta}) \left(\int_0^\ell \frac{dz}{[V(z)]^{\frac{1}{2}}} + \Delta V \int_0^\ell \frac{[(V'(z)/V'(0)) - 1]\, dz}{2[V(z)]^{\frac{3}{2}}} \right. \\ &\qquad \left. + \frac{\Delta V}{V'(0)[V'(\ell)]^{\frac{1}{2}}} \right) \end{aligned} \tag{f6}$$

Whence

$$\tau' = \frac{\Delta\tau}{\Delta V} = (2/\sqrt{2\eta}) \left(\frac{1}{V'(0)[V'(\ell)]^{\frac{1}{2}}} + \int_0^\ell \frac{[(V'(z)/V'(0)) - 1]\, dz}{2[V(z)]^{\frac{3}{2}}} \right. \tag{f7}$$

In computing F it should be noted that by definition the gap voltage pre-

viously referred to as V_0 is

$$V_0 = V(\ell) \tag{f8}$$

In the notation as it has been modified, the transit time is

$$\tau_0 = (2/\sqrt{2\eta}) \int_0^{\ell} \frac{dz}{(V(z))^{\frac{1}{2}}} . \tag{f9}$$

For a constant retarding field in the drift space of magnitude E_0, we can write

$$\eta E_0(\tau_1/2) = v = \sqrt{2\eta V} \tag{f10}$$

Here V is the total energy with which electrons are shot into the drift space. From (f10)

$$\tau_1 = \frac{2\sqrt{2\eta V}}{\eta E_0} \tag{f11}$$

$$\tau_1' = d\tau/dV = \frac{2\sqrt{2\eta V}}{\eta E_0}\left(\frac{1}{2V}\right) = \frac{\tau_0}{2V} . \tag{f12}$$

In the notation used earlier this is

$$\tau_1' = \frac{\tau_0}{2V(\ell)} . \tag{f13}$$

Now we will compare τ' from (f7) with τ_1', the rate of change of transit time for a linear field, taking τ_1' for the same resonator voltage $V(\ell)$ and the same transit time, given by (f1), as the nonlinear field. The factor F relating τ' and τ_1' will then be

$$F = \tau'/\tau_1'$$

$$= 2V(\ell)\left[\left\{\frac{1}{V'(0)[V(\ell)]^{\frac{1}{2}}} + \int_0^{\ell} \frac{[V'(z)/V'(0) - 1]\, dz}{2[V(z)]^{\frac{3}{2}}}\right\} \int_0^{\ell} \frac{dz}{[V(z)]^{\frac{1}{2}}}\right]^{-1} . \tag{f14}$$

If an electron is shot into the drift space with more than average energy, its greater penetration causes it to take longer to return, but it covers any element of distance in less time. Consider a case in which the gradient of the potential is small near the zero potential (much change in penetration for a given change in energy) and larger near the gap. The first term in the brackets of (f14) will be large, and the second is in this case positive. This means that in this type of field the increased penetration per unit energy and the effect of covering a given distance in less time with increased energy work together to give more drift action than in a constant field. However,

we might have the gradient near the gap less than that at the zero potential. In that case the second term in the brackets would be negative. This means a diminution in drift action because the penetration changes little with energy while the electron travels faster over the distance it has to cover.

To show how large this effect of weakening the field near the zero potential point may be, we will consider a specific potential variation, one which approximates the field in a long hollow tubular repeller. The field considered will be that in which

$$V(z) = (e^z - 1)/e^\ell \tag{f15}$$

We obtain

$$F = (e^\ell - 1) + \frac{(e^\ell - 1)^{\frac{1}{2}}}{\tan^{-1}(e^\ell - 1)^{\frac{1}{2}}}. \tag{f16}$$

Now

$$V'(0) = e^{-\ell} \tag{f17}$$

$$V'(z) = 1 \tag{f18}$$

Hence

$$F = \left(\frac{1}{V'(0)} - 1\right) + \frac{\left(\frac{1}{V'(0)} - 1\right)^{\frac{1}{2}}}{\tan^{-1}\left(\frac{1}{V'(0)} - 1\right)^{\frac{1}{2}}}. \tag{f19}$$

This shows clearly how the effective drift angle is increased as the field at the zero potential point is weakened. For instance, if $V'(0) = \frac{1}{2}$, so that the field at the zero potential point is $\frac{1}{2}$ that at the gap, the drift effectiveness for a given number of cycles drift is more than doubled ($F = 2.27$).

There is another approach which is important in that it relates the variations of drift time obtained by varying various voltages. Suppose the gap voltage with respect to the cathode is V_0 and the repeller voltage with respect to the cathode is $-V_R$. Now suppose V_0 and V_R are increased by a factor α, so that the resonator and repeller voltages become αV_0 and $-\alpha V_R$. The zero voltage point at which the electrons are turned back will be at the same position and so the electrons will travel the same distance, but at each point the electrons will go $\alpha^{-\frac{1}{2}}$ times as fast. If, instead of introducing the factor α, we merely consider the voltages V_R and V_0 to be the variables, we see that the transit time can be written in the form

$$\tau = V_0^{-\frac{1}{2}} F(V_R/V_0) \tag{f20}$$

The function $F(V_R/V_0)$ expresses the effect on τ of different penetrations of the electron into the drift field and the factor $V_0^{-\frac{1}{2}}$ tell us that if V_R and V_0 are changed in the same ratio, the drift time changes as one over the square root of either voltage.

We can differentiate, obtaining

$$\partial\tau/\partial V_R = V_0^{-\frac{3}{2}}F'(V_R/V_0) \tag{f21}$$

$$\partial\tau/\partial V_0 = -V_0^{-\frac{3}{2}}((V_R/V_0)F'(V_\tau/V_0) + (1/2)F(V_R/V_0)) \tag{f22}$$

If the electron gains an energy βV in crossing the gap, the effect on τ is the same as if V_0 were increased by βV and V_R were changed by an amount $-\beta V$, because in an acceleration of an electron in crossing the gap the electron gains energy with respect to both the resonator (where the energy is specified by V_0 for an unaccelerated electron) and with respect to the repeller ($-V_R$ for an unaccelerated electron). We may thus write

$$\partial\tau/\partial(\beta V) = \partial\tau/\partial V_0 - \partial\tau/\partial V_R$$

$$= -V_0^{-\frac{3}{2}}[(1 + V_R/V_0)F'(V_R/V_0) + \tfrac{1}{2}F(V_R/V_0)] \tag{f23}$$

This expression (f23) is for the same quantity as (f7). V_0 of (f23) is $V(\ell)$ of (f7). Expressions (f21), (f22) and (f23) compare the effects on drift time of changing the repeller voltage alone, as in electronic tuning, the resonator voltage alone, and of accelerating the electrons in crossing the gap. As making the repeller more negative always decreases the drift time, we see that the two terms of (f22) subtract, and usually $|\partial\tau/\partial V_0|$ will be less than $|\partial\tau/\partial V_R|$. In fact, for a linear variation potential in the drift space and for $V_0 = V_R$, $\partial\tau/\partial V_0 = 0$. Weak fields at the zero potential point make the absolute value of $F'(V_R/V_0)$ larger and hence tend to make both $|\partial\tau/\partial V_R|$ and $|\partial\tau/\partial(\beta V)|$ larger. However, these quantities are not changed in quite the same way.

The reader should be warned that (f14) and (f20)–(f23) apply only for fields not affected by the space charge of the electron beam. For instance, suppose we had a gap with a flat grid and a parallel plane repeller a long way off at zero potential. If edge effects and thermal velocities were neglected, we would have a Child's law discharge. The potential would be zero beyond a certain distance from the repeller, and we would have

$$V(z) = Az^{\frac{4}{3}}$$

According to (f14), F should be infinite. There is no reason to expect infinite drift action, however, for the drift field, which is affected by the fluctuating electron density in the beam, is a function of time, and (f14) does not apply.

APPENDIX VII

Ideal Drift Field

The behavior of reflex oscillators has been analyzed on the basis of a uniform field in the drift space. It can be shown that this is not the drift field which gives maximum efficiency. The field which does give maximum efficiency under certain assumptions is described in this appendix.

Consider a reflex oscillator in which a voltage V appears across the gap. This voltage causes an energy change of $\beta V \cos \theta_1$ for the electron crossing the gap. Here θ_1 is the phase at which a given electron crosses the gap for the first time. The effect of the drift space is to cause the electron to return after an interval τ_a where τ_a is a function of this energy.

$$\tau_a = f(\beta V \cos \theta_1) \tag{g1}$$

Thus, each value of τ_a will occur twice every cycle (2π variation of θ_1). We will have

$$\theta_1 = \omega t_1 \tag{g2}$$

$$\begin{aligned} \theta_a &= \omega(t_1 + \tau_a) \\ &= \theta_1 + \varphi(\theta_1) \end{aligned} \tag{g3}$$

$$\varphi(\theta_1) = \omega \tau_a \tag{g4}$$

Here t_1 is the time at which an electron first crosses the gap and $(t_1 + \tau_a)$ is the time of return to the gap. θ_1 and θ_a are the phase angles of the voltage at first crossing and return.

The net work done by an electron in the two crossings is

$$W = \beta V e[-\cos \theta_1 + \cos (\theta_1 + \varphi(\theta_1))] \tag{g5}$$

If the beam current has a steady value I_0, the power produced will be

$$P = (\beta V I_0/2\pi) \int_{-\pi}^{\pi} [-\cos \theta_1 + \cos (\theta_1 + \varphi(\theta_1))]\, d\theta_1 . \tag{g6}$$

The integral of $\cos \theta_1$ is of course zero. Further, from (g4) we see that

$$\varphi(\theta_1) = \varphi(-\theta_1)$$

Hence

$$\begin{aligned} P &= (\beta V I_0/2\pi) \int_0^{\pi} [\cos (-\theta_1 + \varphi(\theta_1)) + \cos (\theta_1 + \varphi(\theta_1))]\, d\theta_1 \\ &= (\beta V I_0/2\pi) \int_0^{\pi} \cos \varphi(\theta_1) \cos \theta_1\, d\theta_1 . \end{aligned} \tag{g7}$$

As $\cos \varphi(\theta_1)$ cannot be greater than unity, it is obvious that this will have its greatest value if the following holds

$$0 < \theta_1 < \pi/2, \qquad \varphi(\theta_1) = 2n\pi, \qquad \cos \varphi(\theta_1) = +1 \tag{g8}$$

$$\pi/2 < \theta_1 < \pi, \qquad \varphi(\theta_1) = (2m + 1)\pi, \qquad \cos \varphi(\theta_1) = -1 \tag{g9}$$

These conditions are such that for a positive value of $\cos \varphi$ the gap voltage is accelerating giving a longer drift time than obtains for a negative value of $\cos \varphi(\theta_1)$ for which a retarding gap voltage is required. Thus, physically we must have

$$2n > 2m + 1 \tag{g10}$$

The simplest case is that for $n = 1$ and $m = 0$, so that in terms of the gap voltage

$$\begin{aligned} v < 0, \qquad & \varphi(\theta_1) = \pi \\ v > 0, \qquad & \varphi(\theta_1) = 2\pi \end{aligned} \tag{g11}$$

This sort of drift action is illustrated by the curve shown in Fig. 136.

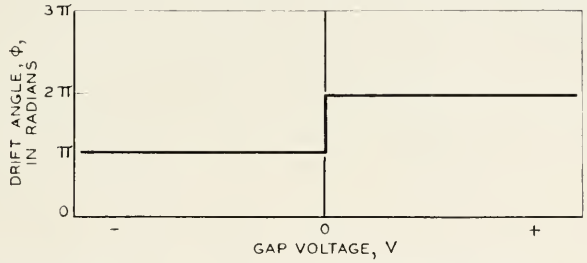


Fig. 136.—Ideal variation of drift time in the repeller region with resonator gap voltage.

The problem of finding the variation with distance which would give this result was referred to Dr. L. A. MacColl who gave the following solution:

Suppose V_0 is the voltage of the gap with respect to the cathode and Φ is the potential in the drift space. Let

$$x_0 = \sqrt{2\eta V_0}/\omega \tag{g12}$$

Here ω is the operating radian frequency. Let x be a measure of distance in the drift field.

$$\begin{aligned} \Phi &= V_0[1 - (x/x_0)^2], \qquad 0 < x < x_0 \\ \Phi &= V_0\{1 - [(x/x_0)^2 + 1]^2/4(x/x_0)^2\}, \qquad x > x_0 \end{aligned} \tag{g13}$$

This potential distribution is plotted in Fig. 137.

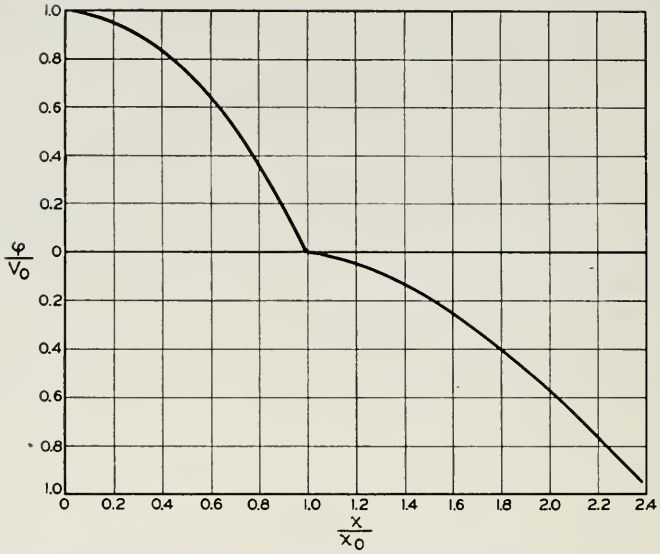


Fig. 137.—Variation of potential in the repeller region vs distance to give the characteristics shown in Fig. 7.1.

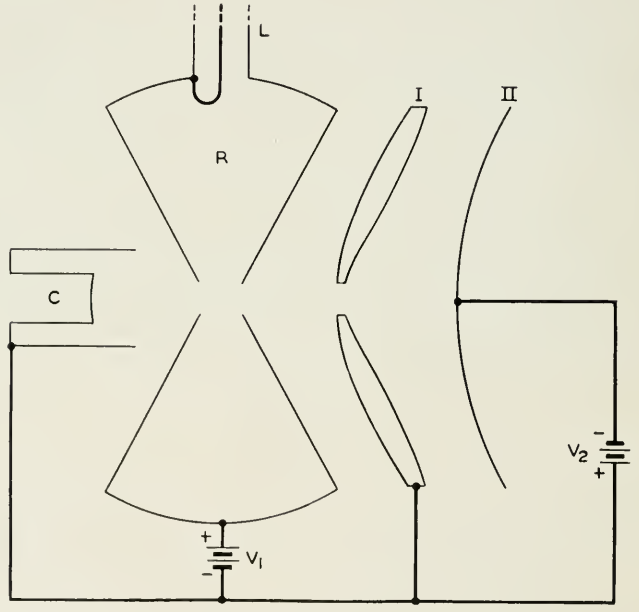


Fig. 138.—Electrodes to achieve approximately the potential variation in the drift region shown in Fig. 7.2.

The shapes for electrodes to realize this field may be obtained analytically by known means or experimentally by measurements in a water tank. The general appearance of such electrodes and their embodiment in a reflex oscillator are shown in Fig. 138. Here C is the thermionic cathode forming part of an electron gun which shoots an electron beam through the apertures or gap in a resonator R. The beam is then reflected in the drift field formed by the resonator wall, zero potential electrode I and negative electrode II, which give substantially the axial potential distribution shown in Fig. 137. Small apertures in the resonator wall and in electrode I allow passage of the electron beam without seriously distorting the drift field. Voltage sources V_1 and V_2 maintain the electrodes at proper potentials. Either suitable convergence of the electron beam passing through the resonator from the gun or axial magnetic focusing will assure return of reflected electrons through the resonator aperture. In addition, the aperture in electrode I forms a converging lens which tends to offset the diverging action of the fields existing between the resonator wall and I, and between I and II. R-f power is derived from resonator R by a coupling loop and line L.

APPENDIX VIII

Electronic Gap Loading

If a measurement is made of gap admittance in the presence and in the absence of the electron beam passing across it once, it will be found that the electron stream gives rise to an admittance component Y. The susceptance is unimportant, but the conductance G can have a noticeable effect on the efficiency of an oscillator.

Petrie, Strachey and Wallis have provided an important expression for this gap conductance due to longitudinal fields when the r-f voltage is small compared with the beam voltage V_0.[25] In this analysis it is presumed that the fields in the beam are due to the voltages on the electrodes only and not to the space change in the beam.[26] This analysis is of such importance that it is of interest to reproduce it in a slightly modified form. We will first consider the general cases of interaction with longitudinal fields and will then consider transverse fields also.

A. *Longitudinal Field*

Assume a stream of electrons flowing in the positive x direction, constituting a current $-I_0$, bunched to have an a-c convection current component

[25] These expressions were communicated to the writers through unpublished but widely circulated material by D. P. R. Petrie, C. Strachey and P. J. Wallis of Standard Telephones and Cables Valve Laboratory.

[26] The expressions are valid in the presence of space charge, but as the field is not known, they cannot be evaluated.

i_1. Now if t_1 is the time a particle passes x_1 and t_2 the time the particle passes x_2, the convection current at x_2 will be

$$i_2 = (-I_0 + i_1)\frac{dt_1}{dt_2} + I_0 . \tag{h1}$$

This merely states that the charge which passes x_1 in the time interval dt_1 will pass x_2 in the time interval dt_2. Suppose the electrons are accelerated at x_1 by a voltage

$$Ve^{j\omega t_1}$$

If V_0 is the voltage specifying the average speed of the electrons, the velocity will be

$$v = (2\eta V_0)^{\frac{1}{2}}(1 + (V/V_0)e^{j\omega t_1})^{\frac{1}{2}} \tag{h2}$$

We then have

$$\begin{aligned} t_2 &= t_1 + \tau(1 + (V/V_0)e^{j\omega t_1})^{-\frac{1}{2}} \\ \tau &= x(2\eta V_0)^{-\frac{1}{2}} \end{aligned} \tag{h3}$$

$$\frac{dt_2}{dt_1} = 1 - \frac{j\omega\tau(V/V_0)e^{j\omega t_1}}{2(1 + (V/V_0)e^{j\omega t_1})} . \tag{h4}$$

Now assume $(V/V_0) \ll 1$. If we neglect higher powers than the first we can replace t_1 by

$$t_1 = t_2 - \tau$$

and obtain

$$\frac{dt_1}{dt_2} = 1 + \frac{j\tau V e^{j\omega(t_1-\tau)}}{2V_0} \tag{h5}$$

and from (h1), neglecting products of two a-c quantities

$$i_2 = i_1 - \frac{j\omega\tau I_0}{2V_0} Ve^{j\omega(t_1-\tau)} . \tag{h6}$$

Suppose we consider an electron stream travelling through a longitudinal field of potential fluctuating as $e^{j\omega t}$ and of magnitude $V(x)$, where $V(x)$ may be complex. Then the current at x_2 due to the action of the field at x_1 is, omitting for convenience the factor $e^{j\omega t}$,

$$di_2 = \frac{jI_0 V'(x_1)}{2V_0} \gamma(x_2 - x_1)e^{-j\gamma(x_2-x_1)}\, dx_1 \tag{h7}$$

$$\gamma = \omega/u_0 \tag{h8}$$

$$u_0 = (2\eta V_0)^{\frac{1}{2}} . \tag{h9}$$

Hence, the convection current at x_2 is

$$i_2 = \frac{-jI_0}{2V_0} \int_{-\infty}^{x_2} V'(x_1)\gamma(x_2 - x_1)e^{-j\gamma(x_2-x_1)}\, dx_1 . \tag{h10}$$

The power flow *from the electron stream to the circuit* is

$$P = \tfrac{1}{2} \int_{-\infty}^{\infty} V'(x_2)i_2^*\, dx_2 \tag{h11}$$

$$P = \frac{jI_0}{4V_0} \int_{-\infty}^{\infty} \int_{-\infty}^{x_2} V'(x_2)V'^*(x_1)\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\, dx_1\, dx_2 . \tag{h12}$$

We have in (h12) an expression, based on the physics of the picture, for power flow from the electron stream to the circuit. This expression is a product of the electron convection current, due to bunching, and the electric field, and the product is integrated from $x = -\infty$ to $x = +\infty$, which is merely a way of including all the a-c fields present. We could as well have integrated between two points a and b between which all a-c fields lie.

We will now go through some strictly mathematical manipulations of (h12), designed to transform it to a more handy form. In the following steps we may regard x_1 and x_2 as merely two different variables of integration, disregarding completely their physical significance.

Suppose in the $x_1\, x_2$ plane, x_1 is measured $+$ to the right and x_2, $+$ upwards. Then the plane is divided into two portions by the line $x_2 = x_1$, and we are integrating over the upper left portion. If we reverse the order of integration we obtain

$$P = \frac{jI_0}{4V_0} \int_{-\infty}^{\infty} \int_{x_1}^{\infty} V'(x_2)V'^*(x_1)\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\, dx_2\, dx_1 . \tag{h13}$$

Let us (a) interchange the variables of integration x_2 and x_1 and (b) take the conjugate of P. We obtain

$$P^* = \frac{jI_0}{4V_0} \int_{-\infty}^{\infty} \int_{x_2}^{\infty} V'(x_2)V'^*(x_1)\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\, dx_1\, dx_2 . \tag{h14}$$

The real part of P is $\frac{1}{2}(P + P^*)$; hence

$$\text{Real} = \frac{jI_0}{8V_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} V'(x_2)V'^*(x_1)\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\, dx_1\, dx_2 . \tag{h15}$$

Let us consider the quantity

$$A = \int_{-\infty}^{\infty} V'(x)e^{j\gamma x}\, dx \tag{h16}$$

$$|A|^2 = AA^* = \left(\int_{-\infty}^{\infty} V'(x_2)e^{j\gamma x_2}\,dx_2\right)\left(\int_{-\infty}^{\infty} V'^*(x_1)e^{-j\gamma x_1}\,dx_1\right)$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V'(x_2)V'^*(x_1)e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2 \tag{h17}$$

$$\frac{\partial |A|^2}{\partial \gamma} = \frac{j}{\beta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V'(x_2)V'^*(x_1)\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2\,. \tag{h18}$$

Hence, we see

$$\text{Real} = \frac{I_0}{8V_0}\,\gamma\,\frac{\partial |A|^2}{\partial \gamma}\,. \tag{h19}$$

B. *Transverse Field*

Suppose we consider the additional power transfer because of deflections. There will be two sources of energy transfer. First, imagine a fluctuating y component of velocity, $\dot{y}$. Let u_0 be the x component of velocity and $-I_0$ the convection current to the right. In a distance dx this will flow against the potential gradient in the y direction a distance

$$dy = (\dot{y}/u_0)dx \tag{h20}$$

and the power flowing to the field from the beam will be

$$dP = -\frac{I_0}{2}\frac{\partial V}{\partial y}\,(\dot{y}^*/u_0)\,dx. \tag{h21}$$

This is not the total power transfer, however. The beam will also suffer a displacement y in the y direction. Now the x component of field varies with displacement; hence the beam will encounter a varying field. We can write the instantaneous power transferred from the beam to the field. Let $(V)_i$ be the instantaneous value of V and $(y)_i$ be the instantaneous value of y. The instantaneous power will be

$$dp = -\left(\frac{d(V)_i}{dx} + \frac{\partial^2(V)_i}{\partial x \partial y}\,(y)_i\right)dx. \tag{h22}$$

Let us compare this with the instantaneous power transferred from the beam to the field by a fluctuating convection current $(i)_i$

$$dp = \frac{\partial(V)_i}{\partial x}\,(i)_i\,. \tag{h23}$$

We see that according to our convention that

$$P = VI^* \tag{h24}$$

we may write from (h22)

$$P = -\frac{I_0^*}{2}\int_{-\infty}^{\infty}\left(\frac{\partial V}{\partial y}\right)_2 dx_2 \int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_1^* dx_1 \tag{h25}$$

The y gradient of the potential at x_1 produces a velocity at x_2

$$\dot{y}_2 = \frac{\eta}{u_0}\int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_1 e^{-j\gamma(x_2-x_1)}\,dx_1 . \tag{h26}$$

It produces a displacement at x_2

$$y_2 = \frac{\eta}{u_0^2}\int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_1 (x_2 - x_1)e^{-j\gamma(x_2-x_1)}\,dx_1 . \tag{h27}$$

Writing the total power as

$$P = P_1 + P_2 . \tag{h28}$$

We have the two contributions from (h22) and (h23) using (h9)

$$P_1 = \frac{-I_0}{4V_0}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^* e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2 \tag{h29}$$

and

$$P_2 = \frac{-I_0}{4V_0}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2}\left(\frac{\partial^2 V}{\partial x\,\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^* (x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2 . \tag{h30}$$

Again, we will turn to mathematical manipulation disregarding the physical significance of the variables. If we change the order of integration, (h30) becomes

$$P_2 = \frac{-I_0}{4V_0}\int_{-\infty}^{\infty}\int_{x_1}^{\infty}\left(\frac{\partial^2 V}{\partial x\,\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1 (x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_2\,dx_1 . \tag{h31}$$

Integrating with respect to x_2 by parts we obtain

$$\begin{aligned} P_2 = -\frac{I_0}{4V_0}\Bigg(&\int_{-\infty}^{\infty}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^* (x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_1\Bigg|_{x_1}^{\infty} \\ &- j\int_{-\infty}^{\infty}\int_{x_1}^{\infty}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^* \gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_2\,dx_1 \\ &- \int_{-\infty}^{\infty}\int_{x_1}^{\infty}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^* e^{j\gamma(x_2-x_1)}\,dx_2\,dx_1\Bigg). \end{aligned} \tag{h32}$$

The first term is zero because $\left(\frac{\partial V}{\partial y}\right)^*$ is zero at $x_1 = -\infty$ and $(x_2 - x_1)$ is

zero at $x_2 = x_1$. Changing the order of integration, we obtain

$$P = +j\frac{I_0}{4V_0}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^*\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2 + \frac{I_0}{4V_0}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^* e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2\,. \tag{h33}$$

From (h29) and (h30) we see that the total power is

$$P = j\frac{I_0}{4V_0}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2}\left(\frac{\partial V}{\partial y}\right)_2\left(\frac{\partial V}{\partial y}\right)_1^*\gamma(x_2 - x_1)e^{j\gamma(x_2-x_1)}\,dx_1\,dx_2. \tag{h34}$$

By the same means resorted to in connection with (h12) we find

$$\text{Real} = \frac{I_0}{8V_0}\gamma\frac{\partial\,|\,B\,|^2}{\partial\gamma} \tag{h35}$$

$$B = \int_{-\infty}^{\infty}\frac{\partial V}{\partial y}e^{j\gamma x}\,dx. \tag{h36}$$

We can go a step further. We have

$$A = \int_{-\infty}^{\infty}\frac{\partial V}{\partial x}e^{j\gamma x}\,dx. \tag{h37}$$

Now

$$\frac{\partial A}{\partial y} = \int_{-\infty}^{\infty}\frac{\partial^2 V}{\partial x\partial y}e^{j\gamma x}\,dx. \tag{h38}$$

Integrating by parts

$$\frac{\partial A}{\partial y} = \frac{\partial V}{\partial y}e^{j\gamma x}\Big|_{-\infty}^{\infty} - j\gamma\int_{-\infty}^{\infty}\frac{\partial V}{\partial y}e^{j\gamma x}\,dx. \tag{h39}$$

The first term is zero, and we see that

$$|\,B\,|^2 = (|\,\partial A/\partial y\,|^2/\gamma^2) \tag{h40}$$

$$\text{Real} = \frac{I_0}{8V_0}\gamma\frac{\partial(|\,\partial A/\partial y\,|^2/\gamma^2)}{\partial\gamma}. \tag{h41}$$

C. *Electronic Gap Loading*

In (h19) and (h41) we have expressed the power flow from the electron stream to the circuit in rather general terms. What we want immediately is the quantity (conductance) giving the power flow from the circuit to the stream for a single gap. Assume we have a single gap with unit peak r-f voltage across it. The power absorbed by the electron stream can be

attributed to a shunt conductance such that

$$P_R = GV^2/2 = G\frac{2I_1(\gamma r_0)}{\gamma r_0 I_0(\gamma r_0)}\left[1 - \frac{I_1(\gamma r_0)}{I_0(\gamma r_0)}\right] \tag{h42}$$

$$G = 2P_R . \tag{h53}$$

Also, in this case $|A|$ is simply β, the modulation coefficient. Hence, the conductance due to action of the longitudinal fields is, from (h18), simply

$$G_1 = \frac{-I_0\gamma}{4V_0}\frac{\partial\beta^2}{\partial\gamma}. \tag{h43}$$

And, due to the action of transverse fields there is another conductance, from (h41)

$$G_2 = \frac{I_0\gamma}{4V_0}\frac{\partial}{\partial\gamma}\left(\frac{1}{\gamma^2}\left(\frac{\partial\beta}{\partial y}\right)^2\right) \tag{h44}$$

$$G = G_1 + G_2 . \tag{h45}$$

These are surprisingly simple and very useful relations.

It is interesting to take an example which will indicate both effects. We have from (b24) for tubes of radius r_0 with a narrow gap between them

$$\beta_s^2 = \beta_r^2[1 - I_1^2(\gamma r_0)/I_0^2(\gamma r_0)]. \tag{h46}$$

Accordingly, the part of the conductance due to the longitudinal field is

$$G_L = F_L(\gamma r_0)I_0/4V_0 \tag{h47}$$

$$\begin{aligned} F_L(\gamma r_0) &= -\gamma\partial\beta_s^2/\partial\gamma \\ &= -2\left[\left(\frac{I_1(\gamma r_0)}{I_0(\gamma r_0)}\right)^2 - \gamma r_0\left(\frac{I_1(\gamma r_0)}{I_0(\gamma r_0)}\right) - \left(\frac{I_1(\gamma r_0)}{I_0(\gamma r_0)}\right)^3\right]. \end{aligned} \tag{h48}$$

Similarly, the part of the conductance due to the transverse field is

$$G_T = F_T(\gamma r_0)I_0/4V_0 \tag{h49}$$

$$F_T(\gamma r_0) = -\gamma\frac{\partial}{\partial\gamma}\frac{1}{r_0^2}\int_0^{r_0}\left(\frac{1}{\gamma}\frac{\partial}{\partial r}\beta_r\right)^2 2r\,dr. \tag{h50}$$

From (b16) we obtain

$$\beta_r = I_0(\gamma r)/I_0(\gamma r_0) \tag{h51}$$

$$\begin{aligned} \frac{1}{\gamma}\frac{\partial\beta_r}{\partial r} &= I_1(\gamma r)/I_0(\gamma r_0)\,\frac{1}{r_0^2}\int_0^{r_0}\left(\frac{1}{\gamma}\frac{\partial}{\partial r}\beta r\right)^2 2r\,dr \\ &= \left[\frac{I_2(\gamma r_0)}{I_0(\gamma r_0)} - \frac{I_1^2(\gamma r_0)}{I_0^2(\gamma r_0)}\right]. \end{aligned} \tag{h52}$$

And from this we obtain

$$F(\gamma r_0) = -2\left[1 - \frac{2I_1(\gamma r_0)}{\gamma r_0 I_0(\gamma r_0)}\right] + 2\left[\frac{I_1(\gamma r_0)}{I_0(\gamma r_0)}\right]^2 - \left[\frac{I_1(\gamma r_0)}{I_0(\gamma r_0)}\right]^3. \tag{h53}$$

The total conductance is

$$G = G_1 + G_2 = (F_L(\gamma r_0) + F_T(\gamma r_0))I_0/4V_0. \tag{h54}$$

In Fig. 139, (G_1V_0/I_0), (G_2V_0/I_0) and (GV_0/I_0) are plotted vs γr_0. It may be seen that while the conductance due to transverse fields may be negative, the total conductance is always positive.

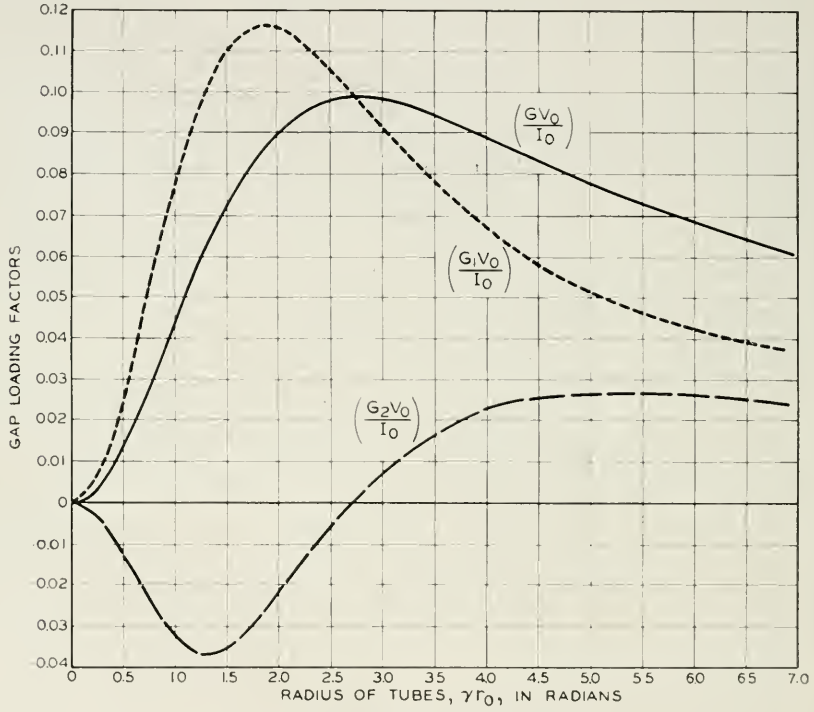


Fig. 139.—Gap loading factor vs radius of tubes forming gap measured in radians. The tubes are supposed to be filled with uniform electron flow. The curve involving G_1 is that for longitudinal effects, that involving G_2 is for transverse effects, and that involving G is for both combined.

This example tends to exaggerate the effects of transverse fields because the beam is assumed to fill the whole tube. In an actual case the beam

would probably not fill the whole tube, and the effect of transverse fields would be less.

Perhaps a more useful expression is one involving longitudinal fields only; that for infinitely fine parallel grids. In this case, if the separation is ℓ

$$\beta^2 = \sin^2(\gamma\ell/2)/(\gamma\ell/2)^2 \tag{h55}$$

and, from (h43)

$$G = \frac{I_0}{2V_0}\frac{\sin(\gamma\ell/2)}{(\gamma\ell/2)}\left[\frac{\sin(\gamma\ell/2)}{(\gamma\ell/2)} - \cos(\gamma\ell/2)\right]. \tag{h56}$$

In Fig. 140, (GV_0/I_0) is plotted vs $(\gamma\ell/2)$. The negative conductance region beyond $\gamma\ell = 2\pi$, familiar through Llewellyn's work with diode oscillators, is of less interest in connection with reflex oscillators.

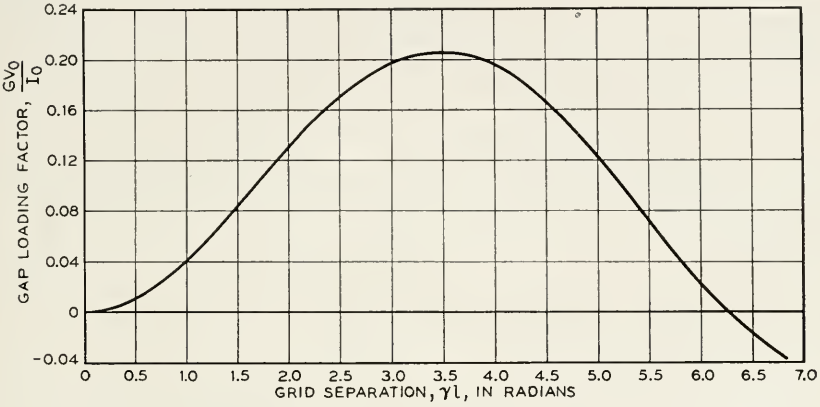


Fig. 140.—Gap loading factor for fine parallel grids vs grid separation in radians.

It is of some interest to compare the electronic gap loading with the small signal electronic conductance due to drift action. Assume, for instance, we have fine parallel plane grids for which $\gamma\ell = \pi$. From (h2) we get

$$(GV_0/I_0) = .202.$$

As the current crosses the gap twice we should count the current involved as twice the d-c beam current I_b

$$G = .404\, I_b/V_0\,.$$

From (2.1)

$$\beta = .633$$

$$\beta^2 = .400.$$

If we assume 3.75 cycles of drift, then from (2.4) the magnitude of the small signal electronic admittance is

$$\begin{aligned} y_e &= \beta^2 I_0 \theta / 2V_0 \\ &= 4.71\ I_0/V_0\ . \end{aligned} \tag{h57}$$

Thus, in this example, the gap loading is about 1/10 of the small signal electronic admittance.

D. *Bunching in the Gap*

An unbunched stream will become bunched due to a single transit across an excited gap. Expression (h10) gives us a means for calculating the extent of this bunching. As an example, we will consider the case of fine parallel grids separated by a distance ℓ. Then the gradient is given by

$$V'(x_1) = V/\ell \tag{h58}$$

from $x_1 = 0$ to $x_1 = x_2 = \ell$, and by zero elsewhere. Thus

$$\begin{aligned} i_2 &= \frac{jI_0 V}{2V_0 \ell} \int_0^\ell \gamma(\ell - x_1) e^{-j\gamma(\ell - x_1)}\, dx_1 \\ i_2/V &= (I_0/V_0)(1/2)[1 - (j/\gamma\ell)(1 - e^{j\gamma\ell})]e^{-j\gamma\ell}. \end{aligned} \tag{h59}$$

It should be noted that for large values of $\gamma\ell$

$$i_2/V = (1/2)(I_0/V_0)e^{-j\gamma\ell}. \tag{h60}$$

For our previous example, if $\gamma\ell = \pi$,

$$i_2/V = .592\ (I_0/V_0)e^{-j3.52}. \tag{h61}$$

If the current is referred to the center of the gap instead of the second grid, we obtain a current i such that

$$i/V = .592\ (I_0/V_0)e^{-j1.96}. \tag{h62}$$

Now, the electrons constituting current i will drift 3.75 cycles and will return across the gap in the opposite direction. To get the induced circuit current I we take this into account and multiply by β

$$\begin{aligned} I/V &= -\beta(i/V)e^{-j\pi(3.75)} \\ &= -.375(I_0/V_0)e^{-j19\pi}. \end{aligned} \tag{h63}$$

This is very nearly of the same size as the electronic gap loading.

The general conclusion seems to be that for typical conditions encountered in reflex oscillators, gap loading and bunching in the gap are small and probably less important than various errors in the theory.

APPENDIX IX

Losses in Grids

Although the general problem of resonator loss calculation is not treated in this paper, grids seem peculiar to vacuum tubes and losses in grids will be discussed briefly.

Assume that we have a pair of grids of some mesh or network material, parallel, circular, and of a diameter compared with the wavelength small enough so that variation of voltage over the grids may be neglected. The capacitance inside of a radius r is

$$\begin{aligned} C &= \epsilon\pi r^2/d \\ \epsilon &= 8.85 \times 10^{-14} \text{ farads/cm.} \end{aligned} \tag{i1}$$

Here d is the separation between the grids. For unit r.m.s. voltage across the grids, the power dissipated in the part of both grids lying in the range dr at r is

$$\begin{aligned} dP &= 2(\omega C)^2(R\,dr/2\pi r) \\ &= \omega^2\epsilon^2\pi R r^3\,dr/d^2. \end{aligned} \tag{i2}$$

Here R is the surface resistivity. The total power is equal to V^2G, where G is the conductance measured at the edge of the grids, and is

$$\begin{aligned} G = P &= \frac{\omega^2\epsilon^2\pi R}{d^2}\int_0^{D/2} r^3\,dr \\ &= \omega^2\epsilon^2\pi R D^4/64\,d^2. \end{aligned} \tag{i3}$$

It is interesting to express ω in terms of λ, the wavelength, c the velocity of light, and then to put in numerical values

$$G = (1/16)\pi^3 c^2\epsilon^2 R D^4/\lambda^2\,d^2 \tag{i4}$$

$$G = 1.39 \times 10^{-5} R D^4/\lambda^2\,d^2. \tag{i5}$$

Now for the copper, the surface resistivity is

$$R_c = .045/\sqrt{\lambda} \tag{i6}$$

where λ is measured in cm. Suppose the grid material is non-magnetic and has N times the low frequency resistivity of copper. Then for it, the surface resistivity will be $N^{\frac{1}{2}}$ times that given by (i6). Suppose that the diameter of the grid wires is $2r$ and the distance center to center is a. If current flowed on one half of the wire surface only, the surface resistivity parallel to the wires would be

$$R = N^{\frac{1}{2}} R_c(a/\pi r). \tag{i7}$$

If we use this as the surface resistivity in (i5) we obtain

$$G = 1.99 \times 10^{-7} \left(\frac{a}{r}\right) \sqrt{N/\lambda}/D^4\lambda^2 d^2. \tag{i8}$$

It is a somewhat remarkable fact that if we work out the loss for a pair of grids with surface resistivity R in one direction and infinite resistivity in the other direction (parallel wire grids) we get just twice the conductance given by (i5). Thus, it appears to be roughly true that if we have a given parallel wire grid, adding wires between the original wires and adding wires perpendicular to the original wires should have about the same effect in reducing circuit loss.

APPENDIX X

Starting of Pulsed Reflex Oscillator

If a reflex oscillator is turned on gradually, the voltage from which the oscillations build up is certainly that due to shot noise in the electron stream. However, if the current is turned on as a pulse of short time of rise, it might be that the microwave voltage produced by the high frequency components of the pulse would be larger than the voltage produced by shot noise, and hence that oscillations would build up from the transient produced by the pulse and not from shot noise. This would be important because presumably the voltages produced by the pulse are always the same and related in the same manner to the time of application of the pulse; thus, in buildup from the transient of the pulse there would be no jitter.

In an effort to decide from which voltage the oscillations build up, we will consider Johnson noise and shot noise voltages.

Associated with a mode of oscillation of the resonator there is a mean stored energy kT. If L and C are the effective inductance and capacitance of the mode and $\overline{I^2}$ and $\overline{V^2}$ are the mean square current and voltage

$$kT = \overline{I^2}L/2 + \overline{V^2}C/2. \tag{j1}$$

On the average, half of the energy is in the capacitance and half in the inductance; thus

$$\begin{aligned} \overline{V^2} &= kT/C \\ &= kT\omega_0/M. \end{aligned} \tag{j2}$$

Here M is the characteristic admittance of the mode and ω_0 is the resonant radian frequency.

As an equivalent circuit for the mode we may use conductance G in shunt with an inductance L and a capacitance C. The impressed Johnson noise

current for a bandwidth B is

$$\overline{I_t^2} = 4kTGB. \tag{j3}$$

Suppose the current crossing the gap is $2I_0$ (a current I_0 with two transits). If the current has full shot noise (no space charge reduction of noise) the impressed shot noise current per unit bandwidth will be

$$\begin{aligned} \overline{I_s^2} &= 2e(2I_0)B \\ &= 4eI_0B. \end{aligned} \tag{j4}$$

Thus, we see that the shot noise voltage will be the Johnson noise voltage times the factor

$$\begin{aligned} \overline{I_s^2} &= \frac{e}{kT}\frac{I_0}{G} \\ &= \frac{11{,}600}{T}\frac{I_0}{G}. \end{aligned} \tag{j5}$$

The conductance G is given by

$$G = M/Q.$$

Let us take reasonable values,

$$\begin{aligned} T &= 293^\circ \text{ K } (20^\circ \text{ C}) \\ I_0 &= .2 \text{ amperes} \\ M &= .06 \text{ mhos} \\ Q &= 400 \text{ (loaded } Q\text{)}. \end{aligned}$$

The values of M and Q are about right for the 2K23 pulsed reflex oscillator. We obtain

$$\frac{\overline{I_s^2}}{\overline{I_t^2}} = 5.3 \times 10^4. \tag{j6}$$

Thus, shot noise is very much more important than Johnson noise.

From (j2) and (j5) we see that the shot noise voltage squared may be expressed as

$$\overline{V_s^2} = \frac{e\omega_0 I_0 Q}{M^2}. \tag{j7}$$

In pulsing a reflex oscillator, the voltage and current will be raised simultaneously; however for the sake of simplicity in calculation, we will assume that the voltage is constant and the injected current is zero at $t = 0$, rises

linearly with time to a value I_0 at a time Δt, and remains constant thereafter. There will be a return current across the gap, the negative of the injected current and delayed by the drift time τ with respect to the injected current. In calculating the response to this applied current, loaded Q will be assumed to be infinite; actually, the shunt resistance will be positive when the current is small and negative when the current becomes large enough so that the electronic conductance is larger in magnitude than the circuit conductance. The assumption of zero conductance should, however, give us an idea of the transient which would be effective in starting the oscillator.

If a charge dq is put onto a capacitance $C = M/\omega_0$ forming part of a resonant circuit of frequency ω_0, the subsequent voltage across the circuit will be

$$dV = \frac{\omega_0}{M} e^{j\omega_0 t}\, dq. \tag{j8}$$

We see that for times late enough so that the injected and returned current are both constant, the voltage due to our assumed current will be

$$V = \frac{\omega_0}{M} I_0 \left(\int_0^{\Delta t} \frac{t_0}{\Delta t} e^{j\omega(t-t_0)}\, dt_0 + \int_{\Delta t}^{t} e^{j\omega(t-t_0)}\, dt_0 - \int_{\tau}^{\tau+\Delta t} \frac{t_0 - \tau}{\Delta t} e^{j\omega(t-t_0)}\, dt_0 - \int_{\tau+\Delta t}^{t} e^{j\omega(t-t_0)}\, dt_0 \right). \tag{j9}$$

Integrating, we find

$$V = \left(\frac{I_0}{M \Delta t\, \omega_0} \right) (1 - e^{-j\omega_0 \tau})(e^{-j\omega_0 \Delta t} - 1). \tag{j10}$$

If we have $n + 3/4$ cycle drift,

$$e^{-j\omega_0 \tau} = -j. \tag{j11}$$

The extreme value of $(e^{-j\omega_0 \Delta t} - 1)$ is -2. For this value we would obtain

$$|V|^2 = \frac{2I_0^2}{M^2 \Delta t^2 \omega_0^2}. \tag{j12}$$

From this and (j7) we obtain

$$\frac{\overline{V_s^2}}{|V|^2} = \frac{eQ\Delta t^2 \omega_0^3}{2I_0}. \tag{j13}$$

Taking the values

$$\begin{aligned} e &= 1.59 \times 10^{-19} \text{ coulombs} \\ I_0 &= .2 \text{ amperes} \\ Q &= 400 \text{ (loaded } Q\text{)} \\ \Delta t &= .2 \times 10^{-6} \text{ seconds} \\ \omega_0 &= 25 \times 10^9 \text{ radians/second.} \end{aligned}$$

We obtain

$$\frac{\overline{V_s^2}}{|V|^2} = 99.2. \tag{j14}$$

The indications are that oscillations will built up from shot noise rather than from the high frequency transients induced by the pulse.

The preceding analysis assumes that the full shot noise voltage will appear across the resonator soon enough to override the pulse. The shot noise voltage will reach full amplitude in a time after application of the current of the order of $Q/f_0 = 2\pi Q/\omega_0$. For the values given above

$$2\pi Q/\omega_0 = .05 \times 10^{-6}$$

as this is a considerably smaller time than the .2 microseconds allotted for the buildup of the pulse, the assumption of full shot noise voltage is presumably fairly accurate.

APPENDIX XI

Thermal Tuning

Two extreme conditions may exist

(1) Cooling is by radiation alone

(2) Cooling is by conduction alone.

(1) *Radiation Cooling*

The rate of change of temperature on heating will be given by

$$\frac{dT}{dt} = \frac{1}{C}[P_i - KT^4]. \tag{k1}$$

Where

T is the absolute temperature of the expanding element.
C is the heat capacity of the element.
K is the radiation loss in watts/(degree Kelvin)^4.
P_i is the power input to the tuner in watts.

It is assumed that the temperature of the surroundings is constant and the power radiated to the expanding element is included in P_i. Let

$$P_i = KT_e^4 \quad \text{and} \quad T_r = \frac{T}{T_e}. \tag{k2}$$

T_r is then a reduced temperature since T_e is the temperature which the expanding element would reach at equilibrium for a given power input, i.e. $T_e = \left(\frac{P_i}{K}\right)^{1/4}$. T_r is then always less than 1.

Upon substitution of $(k2)$ in $(k1)$

$$\frac{dT_r}{dt} = \frac{K}{C} T_e^3[1 - T_r^4] \tag{k3}$$

which may be integrated to

$$\frac{C}{2KT_e^3} [\tan^{-1} T_r + \tanh^{-1} T_r] = t + t_0 \tag{k4}$$

where t_0 is a constant of integration. Let

$$F_1(T_r) = [\tan^{-1} T_r + \tanh^{-1} T_r]. \tag{k5}$$

This function is plotted in Fig. 79. In order to determine the cycling time for heating τ_h assume

At	$t = 0,$	$T_r = T_{rc}$	i.e. $T_{rc} = T_c/T_m$
	$t = \tau_h,$	$T_r = T_{rh}$	$T_{rh} = T_h/T_m$

where

T_m is the value of the temperature T_e corresponding to the maximum power input P_m.
T_h is the temperature corresponding to one band limit.
T_c is the temperature corresponding to the other band limit.
$T_h > T_c$.

The cycling time for heating τ_h is then

$$\tau_h = \frac{C}{2KT_m^3} [\tan^{-1} T_{rh} - \tan^{-1} T_{rc} + \tanh^{-1} T_{rh} - \tanh^{-1} T_{rc}] \tag{k6}$$

$$= \frac{C}{2KT_m^3} [F_1(T_{rh}) - F_1(T_{rc})] \tag{k7}$$

which gives the time required for the expanding element to rise in temperature from T_{rc} to T_{rh}, i.e. from T_c to T_h.

If we reduce the power input and wish to determine the cooling time the analysis is similar. If the power input from electron bombardment is reduced to zero there will still be power input to the tuner. The residual power which is kept to the minimum possible level comes from such sources as heat radiated from the cathode and general heating of the envelope by the oscillator section.

Let P_0 be the value of the reduced power input.

Then $$-\frac{dT}{dt} = \frac{1}{C}(KT^4 - P_0) \tag{k8}$$

or $$\frac{dT_s}{dt} = -\frac{K}{C}T_0^3(T_s^4 - 1). \tag{k9}$$

Here $$T_s = \frac{T}{T_0} \quad \text{and} \quad P_0 = KT_0^4 \tag{k10}$$

where P_0 is the power from other sources than direct bombardment. In this case T_s is always greater than 1.
Integration yields

$$\frac{C}{2KT_0^3}[\tan^{-1} T_s + \operatorname{ctnh}^{-1} T_s] = t + t_0 . \tag{k11}$$

Let $$F_2(T_s) = [\tan^{-1}T_s + \operatorname{ctnh}^{-1}T_s]. \tag{k12}$$

This function is plotted in Fig. 80. To determine the cycling time for cooling assume

At time $t = 0$, $T_s = T_{sh}$ i.e. $T_{sh} = T_h/T_0$

$t = \tau_C$, $T_s = T_{sc}$ $T_{sc} = T_c/T_0$.

Then $$\tau_C = \frac{C}{2KT_0^3}[F_2(T_{sc}) - F_2(T_{sh})] \tag{k13}$$

giving the time for the contracting element to cool from temperature T_{sh} to T_{sc}; i.e. from T_h to T_c .

(2) *Conduction Cooling*

The rate of change of temperature on heating will be given by

$$\frac{dT}{dt} = \frac{1}{C}(P_i - kT) \tag{k14}$$

where

T is the temperature difference between the tuning strut and the heat sink.
C is the heat capacity of the strut.
k is the conduction loss in watts/°C.
P_i is the power into the tuner.

The solution of (k14) is then,

$$\frac{P_i}{k} - T = \left(\frac{P_i}{k} - T_0\right)e^{-(k/C)t} \tag{k15}$$

where T_0 is the temperature difference at $t = 0$.

If

the temperature difference from the sink is $T_0 = T_c$ at the cooler limit of the band.

T_h is the temperature difference from the sink at the hotter limit of the band.

P_m is the maximum permitted input, then the cycling time for heating

$$\tau_h = \frac{C}{k} \log_e \frac{P_m - kT_c}{P_m - kT_h}. \tag{k16}$$

On cooling

$$\frac{dT}{dt} = -\frac{k}{C} t \tag{k17}$$

from which

$$T = T_0 e^{-(\kappa/C)t} \tag{k18}$$

$$\tau_c = \frac{C}{k} \log_e \frac{T_h}{T_c}. \tag{k19}$$

Acknowledgments

A large number of people have contributed to the work described in this paper. The program was carried on under the general direction of J. R. Wilson whose support and advice are gratefully acknowledged. J. O. McNally gave more specific direction to these efforts and is responsible for many of the design features. We are greatly indebted to him for his close support and encouragement.

Members of the Bell Laboratories technical staff concerned with various portions of the developments were E. M. Boone, K. Cadmus, R. Hanson, C. A. Hedberg, P. Kusch, C. G. Matland, H. E. Mendenhall, R. M. Ryder and R. C. Winans. R. S. Gormley of the Western Electric Company made considerable contributions to the work on the 2K25.

The mechanical design, construction and specifications for manufacture were carried out by a group of engineers under the direction of V. L. Ronci. This group included D. P. Barry, F. H. Best, S. O. Ekstrand, G. B. Gucker, C. Maggs, W. D. Stratton, R. L. Vance and E. J. Walsh. L. A. Wooten and his associates cooperated in numerous chemical problems throughout this work.

The authors wish to make particular acknowledgement of the work of two technical assistants, Miss Z. Marblestone and F. P. Dreschler, whose

diligent and painstaking efforts were a direct contribution to the successful completion of most of the designs described herein.

The NDRC Radiation Laboratories at the Massachusetts Institute of Technology extended very close cooperation both on problems of design and applications of the tubes. A. G. Hill and J. B. H. Kuper were particularly helpful with technical advice and support. H. V. Neher was responsible for the initial design of the 2K50 and assisted on its final development at the Bell Laboratories. G. Hobart assisted on the 2K50 at M.I.T. and also on the final development as a resident at Bell Laboratories.

Professor R. D. Mindlin on leave from Columbia University served as a consultant on vibration and stress analysis problems particularly on the design of diaphragms and mechanical and thermal tuning mechanisms.

In the preparation of the manuscript the authors are particularly indebted to R. M. Ryder for his extension of the simple bunching theory which is included as a part of this paper. L. A. MacColl provided the solution for several mathematical problems.

Abstracts of Technical Articles by Bell System Authors

Investigation of Oxidation of Copper by Use of Radioactive Cu Tracer.[1] J. Bardeen, W. H. Brattain, and W. Shockley. A very thin layer of radioactive copper was electrolytically deposited on a copper blank. The surface was then oxidized in air at 1000°C for 18 minutes, giving an oxide layer with a thickness of 1.25×10^{-2} cm. After quenching, successive layers of the oxide were removed chemically, and the copper activity in each layer was measured. The observed self-diffusion of radioactive copper in the oxide agrees quantitatively with a theory based on the following assumptions: (a) The oxide grows by diffusion of vacant Cu^{+} sites from the outer surface of the oxide inward to the metal. (b) The concentration of vacant sites as the oxygen-oxide interface is independent of the oxide thickness, and drops linearly from this constant value to zero at the metal boundary. (c) Accompanying the inward flow of vacant sites, there is a flow of positive electron holes such as to maintain electrical neutrality. (d) Self-diffusion of copper ions takes place only by motion into vacant sites. The results give a fairly direct confirmation of the theory of oxidation first suggested by Wagner.

A New Magnetic Material of High Permeability.[2] O. L. Boothby and R. M. Bozorth. This paper describes the preparation, heat treatment, and properties of *supermalloy*, a magnetic alloy of iron, nickel, and molybdenum. In the form of 0.014 in. sheet it has an initial permeability of 50,000 to 150,000, a maximum permeability of 600,000 to 1,200,000, coercive force of 0.002 to 0.005 oersted, and a hysteresis loss of less than 5 ergs/cm³/cycle at B = 5000. Transformer cores made of insulated 0.001 in. tape, spirally wound, have about the same initial permeability and a maximum permeability of 200,000 to 400,000. The alloy has a Curie point of 400°C and appears to have an order-disorder transformation temperature somewhat above 500°C.

Magnetoresistance and Domain Theory of Iron-Nickel Alloys.[3] R. M. Bozorth. Measurements of change of electrical resistivity with magnetization and with tension are reported for iron-nickel alloys containing 40 to 100 per cent nickel. When the magnetostriction is negative (81 to 100 per

[1] *Jour. of Chemical Physics*, December 1946.
[2] *Jour. Applied Physics*, February 1947.
[3] *Phys. Rev.*, Dec. 1 and 15, 1946.

cent nickel), tension (σ) decreases resistivity, and magnetic field (H) increases it. Domain theory predicts the ratio σ/H at which the resistivity is equal to that of the unmagnetized specimen, and the theory is accurately confirmed. Measurements are made in transverse as well as longitudinal magnetic fields, and the difference between the resistances so measured is shown to be independent of the distribution of domains in the unmagnetized state; the erratic results reported in the literature are thus explained and avoided. When magnetostriction is positive, the limiting changes of resistivity with field and tension are sometimes found to be different; this is shown to be caused by the variation of magnetostriction with crystallographic direction.

A Wide-Tuning-Range Microwave Oscillator Tube.[4] JOHN W. CLARK and ARTHUR L. SAMUEL. This paper describes a reflex-type velocity-variation oscillator tube with a wide tuning range in the microwave band. The tube will oscillate from 2000 to 13,000 megacycles, but practical tuning considerations limit the band in any one circuit to a two-to-one frequency range. The problems involved in the design and a description of the various elements are given.

Accelerated Ozone Weathering Test for Rubber.[5] JAMES CRABTREE and A. R. KEMP. Light-energized oxidation and cracking by atmospheric ozone are the agencies chiefly responsible for the deterioration of rubber outdoors. Since these processes are separate and distinct, it is proposed to distinguish between them in the evaluation of rubber for resistance to weathering. An accelerated test for susceptibility to atmospheric ozone cracking is discussed. Apparatus for conducting the test and for measurement of ozone in minute concentration is described in detail.

Measurements in Communications.[6] N. B. FOWLER. For convenient reference, some of the more common measurement units and scales used in communication engineering are presented in tabular form together with supplementary explanatory text. Included in the table, which also indicates the limitations involved, are quantities used in measuring power, volume, circuit noise, sound, light, radio fields, crosstalk coupling, and certain other transmission concepts.

An Improved 200-Mil Push-Pull Density Modulator.[7] J. G. FRAYNE, T. B. CUNNINGHAM and V. PAGLIARULO. A completely new variable-

[4] *Proc. I.R.E. and Waves and Electrons*, January 1947.
[5] *Indus. & Engg. Chemistry, Analytical Edition*, December 1946.
[6] *Electrical Engineering*, February 1947.
[7] *Jour. S.M.P.E.*, December 1946.

density modulator utilizing a three ribbon push-pull valve is described. The entire valve is sealed by the force of the Alnico V permanent magnet on the Permendur pole pieces. Signal is applied to the center ribbon and noise-reduction currents are applied to the outer ribbons. True class A push-pull operation is obtained from the two component single ribbon valves by the use of an inverter prism which aligns the modulating and noise-reduction edges of each aperture.

An anamorphote condenser lens is used to eliminate lamp filament striations at the valve ribbon plane. An anamorphote objective lens gives a 4:1 reduction of the valve aperture in the vertical plane at the film and a 2:1 reduction along the length of the sound track. A meter is supplied to measure exposure as well as setting up "bias." A photocell monitor is supplied and a "blooping" light for indicating synchronous start marks.

Mathematical analysis of the exposure produced by the modulating ribbon is appended as well as a similar analysis of the four ribbon push-pull valve which the new valve supersedes.

Factors Governing the Intelligibility of Speech Sounds.[8] N. R. FRENCH and J. C. STEINBERG. The characteristics of speech, hearing, and noise are discussed in relation to the recognition of speech sounds by the ear. It is shown that the intelligibility of these sounds is related to a quantity called articulation index which can be computed from the intensities of speech and unwanted sounds received by the ear, both as a function of frequency. Relationships developed for this purpose are presented. Results calculated from these relations are compared with the results of tests of the subjective effects on intelligibility of varying the intensity of the received speech, altering its normal intensity-frequency relations and adding noise.

Short Duration Auditory Fatigue as a Method of Classifying Hearing Impairment.[9] MARK B. GARDNER. Earlier studies have classified deafness cases into two general groups, those having functional disorders of the middle ear and those having impairments resulting from atrophy of the nerve fibers terminating along the basilar membrane (conductive and nerve deafness types, respectively). Such classifications have been made using bone conduction threshold measurements and unilateral loudness balance results as the basis for differentiation. Bone conduction results, however, are often subject to considerable error while the unilateral loudness balance technique can only be applied to individuals having one normal and one impaired ear. These limitations introduce a need for a completely independent monaural method of classifying deafness types. This is particularly true for the selec-

[8] *Jour. Acous. Soc. Amer.*, January 1947.
[9] *Jour. Acous. Soc. Amer.*, January 1947.

tion of candidates suitable for the fenestration operation for the restoration of hearing in otosclerosis (immobilized stapes). The present paper is concerned with an investigation of short time auditory fatigue as a method of obtaining an impairment analysis. In this study, it was found that the fatigue of the conductively deafened observer was similar to the normal observer except the onset of fatigue was shifted by the amount of the threshold loss. For the nerve deafened observer, on the other hand, the onset of fatigue was found to occur at normal intensity levels. The occurrence of excessive fatigue in one of the nerve type impairment cases investigated appears to offer additional information on the nature of the lesion.

A Sampling Procedure for Design Tests of Electron Tubes (Sponsored by Joint Electron Tube Engineering Council).[10] S. W. Horrocks, P. M. Dickerson, H. F. Dodge,* E. R. Ott, H. G. Romig,* W. B. Rupp, J. R. Steen, R. E. Warfham, and A. K. Wright. The Committee on Sampling Procedure was established on July 21, 1943 as part of the Electron Tube Section of the Radio Manufacturers Association (RMA). The purpose of the Committee is to develop sampling methods and to act in an advisory capacity towards standardization of Sampling Procedures throughout the Electron Tube industry. This Committee was later embodied as a main Committee of the Joint Electron Tube Engineering Council of the RMA and NEMA. This Council was established in 1945 to handle all engineering matters for the Electron Tube industry for both trade associations, Radio Manufacturers Association and National Electrical Manufacturers Association.

One of the earliest projects handled by the Committee was the development of a statistically sound sampling inspection procedure for so-called "design tests" of electron tubes. In general, design tests relate to characteristics that are normally quite stable and are relatively less important to the consumer. The nature of these tests is such that only relatively small samples are practicable. The Joint Army-Navy Specification JAN-1A incorporated a sampling plan for design tests allowing (1) not more than 10% of the sample tubes to contain design test defects of any one kind and (2) not more than 20% of the sample tubes to contain design test defects of any kind. Because of the extremely wide range in lot sizes for different classes of electron tubes, such a simplified sampling plan was in effect too strict for small lot sizes and too liberal for large lot sizes. Moreover, no distinction was made in the relative seriousness of different kinds of design test defects. The Committee accordingly set about to prepare a sampling inspection plan that would be relatively free of these faults. The new procedure covers all

[10] *Industrial Quality Control*, November 1946.
* *Of Bell Tel. Labs.*

aspects of the acceptance problem and provides an operationally definite criterion for reducing inspection for a product whose quality is regularly well controlled within the intent of the specification.

The procedure developed by the Committee was approved by J.E.T.E.C. (Joint Electron Tube Engineering Council), was approved by the JAN Committee in September 1945, and is reproduced in full in this article. It will be noted that the procedure provides for two Acceptable Quality Levels (AQL), namely 6% defective and 3% defective for individual design test characteristics. Each design test of a particular type of electron tube is classified as either a Standard Design Test with an AQL = 6% or a Special Design Test with an AQL = 3%. For any design test, if product submitted for inspection has quality equal to the AQL, the chances of acceptance are of the order of 94 to 98 out of 100. If quality runs consistently better than the AQL, reduced inspection is permitted, thus serving as an incentive for the manufacturer to strive for better quality. The operating characteristics of the sampling plans involved are appended to this article and show the degree to which the plans will discriminate for various levels for submitted quality.

Resonant Circuit Modulator for Broad Band Acoustic Measurements.[11] Gordon Ferrie Hull, Jr.* A modulation method is described whereby a broad band frequency response is obtained for recording of sound. In particular low frequency sound approaching zero c.p.s. can be recorded. The theory of the resonant circuit modulating principle is first discussed followed by a description of the apparatus which was constructed for this purpose.

Quality Reporting—Putting Inspection Results to Work.[12] Harold R. Kellogg. Quality reporting is an integral part of the general inspection problem. It cannot be divorced from the logic and aims of an overall inspection program. A discussion of quality reporting should therefore include consideration of (1) inspection procedures, including the collection of data; (2) appraisal of the data; (3) reporting and publicizing results. This outlines the program as it is discussed in this paper.

Properties of Monoclinic Crystals.[13] W. P. Mason. Two crystals of the monoclinic sphenoidal class have been found which have modes of vibration with zero temperature coefficients of frequency, and this property together

[11] *Jour. Applied Physics*, December 1946.

* This research was carried out while the author was a member of the Technical Staff of the Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

[12] *Industrial Quality Control*, November 1946.

[13] *Phys. Rev.*, Nov. 1 and 15, 1946.

with the high electromechanical coupling and the high Q's make it appear probable that these crystals may have considerable use as a substitute for quartz which is difficult to obtain in large sizes. These crystals are ethylene diamine tartrate ($C_6H_{14}N_2O_6$) and dipotassium tartrate ($K_2C_4H_4O_6$, $\frac{1}{2}H_2O$). Complete measurements of the elastic, piezoelectric, and dielectric constants of the dipotassium tartrate (DKT) crystal are given in this paper. The crystal has 4 dielectric constants, 8 piezoelectric constants, and 13 elastic constants. A discussion is given in the appendix of the method of measuring these constants by the use of 18 properly oriented crystals.

An Acoustic Constant of Enclosed Spaces Correlatable with Their Apparent Liveness.[14] J. P. Maxfield and W. J. Albersheim. An acoustic constant called liveness is derived, which constant is correlatable with the acoustic properties of the enclosed space and with the distance between the sound source and the listener. This constant represents the ratio of a time integral of the energy density of the reverberant sound to the unintegrated energy density of the direct sound. The validity of this constant is substantiated by empirical data. Certain subjective effects of monaurally reproduced sounds as a function of the liveness of its pick-up conditions are briefly discussed.

Directional Couplers.[15] W. W. Mumford. The directional coupler is a device which samples separately the direct and the reflected waves in a transmission line. A simple theory of its operation is derived. Design data and operating characteristics for a typical unit are presented. Several applications which utilize the directional coupler are discussed.

Theory of the Beam-Type Traveling-Wave Tube.[16] J. R. Pierce. The small-signal theory of the beam traveling-wave tube has been worked out. The equations predict three forward waves, one increasing and two attenuated, and one backward wave which is little affected by the electron stream. The waves are partly electromagnetic and partly disturbance in the electron stream. The dependence of the wave propagation coefficients on voltage, current, circuit loss, and the other properties of the transmission mode which propagates energy and the cut-off transmission modes is given. Expressions for gain and noise figure and an estimate of power output are given. Appendix A gives an expression for the field in a uniform transmission system due to impressed current (as, of an electron stream) in terms of the parameters of the transmission modes. Appendix B calculates the propagation

[14] *Jour. Acous. Soc. Amer.*, January 1947.
[15] *Proc. I.R.E.*, February 1947.
[16] *Proc. I.R.E.*, February 1947.

constant and the field for unit power flow for the gravest mode of helical transmission system.

Traveling-Wave Tubes.[17] J. R. Pierce and L. M. Field. Very-broad-band amplification can be achieved by use of a traveling-wave type of circuit rather than the resonant circuit commonly employed in amplifiers. An amplifier has been built in which an electron beam traveling with about 1/13 the speed of light is shot through a helical transmission line with about the same velocity of propagation. Amplification was obtained over a bandwidth 800 megacycles between 3-decibel points. The gain was 23 decibels at a center-band frequency of 3600 megacycles.

Attenuation of Forced Drainage Effects on Long Uniform Structures.[18] Robert Pope. When forced drainage is applied to an underground metallic structure to provide cathodic protection, the greatest effects on the structure and earth potentials occur in the vicinity of the drainage point and anode. These effects taper off as the distance from the drainage point increases and even in the relatively simple case of a long, uniform structure, the manner in which these effects taper off or attenuate is quite complex. However, by making a few justifiable assumptions, relatively simple equations are developed which provide sufficiently accurate results in most practical cases. Furthermore, the simple equations bring out more clearly the relative importance of the various factors involved than do the more rigorous equations. The approximate equations have been used with fair success in predicting the effects of drainage on underground telephone cables in conduit and on buried coated cables. They should apply quite accurately to coated pipes, and there are examples of reasonably good application on some bare pipes.

The soil and structure characteristics which enter into the equations are discussed, and the units used established.

Alkaline Earth Porcelains Possessing Low Dielectric Loss.[19] M. D. Rigterink and R. O. Grisdale. Alkaline earth porcelains have been prepared from mixtures of clay, flint, and synthetic fluxes consisting of clay calcined with at least three alkaline earth oxides. These porcelains possess excellent dielectric properties, have low coefficients of thermal expansion, are white, and are especially valuable as bases for deposited carbon resistors for which they were developed. Their characteristics make it probable that other uses will be found for materials of this type.

An illustrative composition is 50.0% Florida kaolin, 15.0% flint (325

[17] *Proc. I.R.E.*, February 1947.
[18] *Corrosion*, December 1946.
[19] *Jour. Amer. Ceramic Soc.*, March 1, 1947.

mesh), 35.0% calcine (200 mesh). The composition of the calcine is 40.0% Florida kaolin, 15.0% $MgCO_3$, 15.0% $CaCO_3$, 15.0% $SrCO_3$, 15.0% $BaCO_3$, calcined at 1200°C. The electrical properties of this body at 1 mc. are Q at 25°C, 2160; Q at 250°C, 280; Q at 350°C, 90; specific resistance at 150°C, $10^{13.5}$ ohm-cm. and at 300°C, $10^{10.7}$ ohm-cm.

A Coaxial-Type Water Load and Associated Power-Measuring Apparatus.[20] R. C. Shaw and R. J. Kircher. This paper presents a description of a coaxial-type water load and associated equipment suitable for measuring peak pulse powers of the order of a megawatt. Water-cell loads have been designed to operate at wavelengths of from 10 to 40 centimeters, where the average dissipation is of the order of 300 watts. Ordinary tap water is used in the load to dissipate the radio-frequency power.

The Ammonia Spectrum and Line Shapes Near 1.25-cm Wave-Length.[21] Charles Hard Townes. The ammonia "inversion" lines near 1.25-cm wave-length are resolved, their widths being decreased at low pressures to 200 kilocycles. Line shapes, intensities, and frequencies are measured and correlated with theory. Calculated intensities and Lorentz-type broadening theory fit experimental results if frequency of collision is fifteen times greater than that measured by viscosity methods. Splitting due to rotation is in fair agreement with a recalculation of theoretical values. A saturation effect is observed with increase of power absorbed per molecule and an interpretation made.

Non-Uniform Transmission Lines and Reflection Coefficients.[22] L. R. Walker and N. Wax. A first-order differential equation for the voltage reflection coefficient of a non-uniform line is obtained and it is shown how this equation may be used to calculate the resonant wave-lengths of tapered lines.

Temperature Coefficient of Ultrasonic Velocity in Solutions.[23] G. W. Willard. Extensive measurements have been made, at ten megacycles, of the temperature dependence of ultrasonic velocity in liquids and liquid mixtures. All single liquids tested, except water, were found to have large negative temperature-coefficients in the temperature range of zero to 80°C. Water has a large positive coefficient at room temperature, decreasing to zero at 74°C and then becoming negative (with a peak velocity of 1557

[20] *Proc. I.R.E. and Waves and Electrons*, January 1947.
[21] *Phys. Rev.*, Nov. 1 and 15, 1946.
[22] *Jour. Applied Physics*, December 1946.
[23] *Jour. Acous. Soc. Amer.*, January 1947.

meters/sec). Solutions in water of various other liquids (and of some solids) give parabolic velocity *vs.* temperature curves like that for water but with the peak velocity and peak temperature values shifting with the concentration of the solution. In general increasing the concentration raises the peak velocity slightly and lowers the peak temperature markedly from the values for water alone. It has also been found possible by compounding three-component solutions to adjust the values of the peak velocity and peak temperature independently within a narrow range of velocities and a wide range of temperatures.

Measurements of Ultrasonic Absorption and Velocity in Liquid Mixtures.[24] F. H. Willis. The absorption (α) and velocity (V) of sound in liquid mixtures were measured at four frequencies (ν) in the range 3.8 to 19.2 mc, using the Debye-Sears-Lucas-Biquard optical technique improved by the addition of a differential photoelectric cell indicator. This improvement permitted the use of lower sound intensities together with a wider sound beam than in the visual extinction method, thus improving conditions with respect to cavitation and beam distortion. In the mixtures investigated, α/ν^2 was found to be independent of frequency within the accuracy of the method, and there was no measurable dispersion of acoustic velocity. An absorption peak at intermediate concentrations not shifting with frequency was found in mixtures of acetone and water, and of ethyl alcohol and water, but was not in evidence in mixtures of acetone and ethyl alcohol, and of glycerol and water. The absorption peaks await theoretical explanation.

Measuring Inter-Electrode Capacitances.[25] C. H. Young. New bridge, developed for measurement of extremely small values in high frequency tubes, useful to two-billionths of a microfarad.

[24] *Jour. Acous. Soc. Amer.*, January 1947.
[25] *Tele-Tech*, February 1947.

Contributors to this Issue

William M. Goodall, B.S., California Institute of Technology, 1928; Bell Telephone Laboratories, 1928—. Mr. Goodall has worked on research problems in connection with the ionosphere, radio transmission and early radio relay studies, radar modulators, and more recently microwave radio relay systems.

J. P. Kinzer, M.E., Stevens Institute of Technology, 1925. B.C.E., Brooklyn Polytechnic Institute, 1933. Bell Telephone Laboratories, 1925—. Mr. Kinzer's work has been in the development of carrier telephone repeaters; during the war his attention was directed to investigation of the mathematical problems involved in cavity resonators.

J. R. Pierce, B.S. in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936—. Engaged in study of vacuum tubes.

Allen F. Pomeroy, B.S. in E.E., Brown University, 1929. Public Service Electric and Gas Company, Electrical Testing Laboratory, 1923–1925, 1926; Weston Electrical Instrument Corporation, 1927; Bell Telephone Laboratories, 1929—. Since 1936 Mr. Pomeroy has been principally occupied in developing equipments to measure attenuations, phase shifts, envelope delays, and reflection coefficients for systems suitable for television transmission, and during the war in the development of radar testing equipment.

W. G. Shepherd, B. S. in Electrical Engineering, University of Minnesota, 1933; Ph.D. in Physics, University of Minnesota, 1937. Bell Telephone Laboratories, Inc., 1937—. From 1937 to 1939 Dr. Shepherd was engaged in non-linear circuit research. Since 1939 he has been engaged in the design of electron tubes.

I. G. Wilson, B.S. and M.E., University of Kentucky, 1921. Western Electric Co., Engineering Department, 1921–25. Bell Telephone Laboratories, 1925—. Mr. Wilson has been engaged in the development of amplifiers for broad-band systems. During the war he was project engineer in charge of the design of resonant cavities for radar testing.

VOLUME XXVI OCTOBER, 1947 NO. 4

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Public Library
Kansas City, Mo.

The Radar Receiver.....*L. W. Morrison, Jr.* 693

High-Vacuum Oxide-Cathode Pulse Modulator Tubes
C. E. Fay 818

Polyrod Antennas.....*G. E. Mueller and W. A. Tyrrell* 837

Targets for Microwave Radar Navigation
Sloan D. Robertson 852

Tables of Phase Associated with a Semi-Infinite Unit Slope of Attenuation.....*D. E. Thomas* 870

Abstracts of Technical Articles by Bell System Authors... 900

Contributors to this Issue..... 904

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50¢ per copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORS

R. W. King J. O. Perrine

EDITORIAL BOARD

W. H. Harrison O. E. Buckley
O. B. Blackwell M. J. Kelly
H. S. Osborne A. B. Clark
J. J. Pilliod F. J. Feely

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are 50 cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1947
American Telephone and Telegraph Company

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXVI *October, 1947* *No. 4*

The Radar Receiver

By L. W. MORRISON, JR.

TABLE OF CONTENTS

Introduction ... 694
1. Radar Receiver Design Considerations ... 695
1.1 The Military Radar System ... 695
1.2 The Function of the Radar Receiver ... 696
1.21 Characteristics of the Radar Receiver Input Signal ... 698
1.22 Character of the Output of a Radar Receiver ... 701
1.3 Composition of the Radar Receiver ... 703
2. Radar Receiver Component Design ... 706
2.1 The Radar Receiver Input Circuit ... 706
2.11 Input Signal Characteristics ... 707
2.12 Input Circuit Noise Considerations ... 707
2.13 1000-Mc Radio-Frequency Amplifier Design ... 709
2.14 The Radar Converter ... 712
2.15 The Radar Receiver Beat Oscillator ... 721
2.16 Typical Radar Input Circuit Designs ... 725
2.2 The Radar Intermediate Frequency Amplifier ... 731
2.21 IF Amplifier Requirements ... 731
Band Width ... 731
Gain Characteristics ... 733
Intermediate Midband Frequency ... 734
The Second Detector ... 735
2.22 IF Amplifier Input Circuit Design ... 735
2.23 Interstage Circuit Design ... 739
2.24 Second Detector Design ... 743
2.25 Typical Component Designs ... 743
2.3 The Radar Video Amplifier ... 749
2.31 Gain-Frequency Considerations ... 749
2.32 Gain-Amplitude Considerations ... 751
2.33 D-c Restoration Methods ... 753
2.34 Typical Radar Receiver Video Amplifier Circuits ... 755
2.4 The Radar Indicator ... 757
2.41 Classification of Radar Display Types ... 757
2.42 The Cathode-Ray Tube ... 762
Electrostatic Deflection Type ... 763
Magnetic Deflection Type ... 764
Characteristics of the Fluorescent Screen ... 765
2.43 Typical Radar Indicator Component Designs ... 767
2.5 The Radar Sweep Circuit ... 773
2.51 Function ... 773
2.52 The Timing Wave Form Generator ... 776
The Multivibrator ... 776
Typical Timing Wave Circuits ... 779
2.53 The Sweep Wave Form Generator ... 781
2.54 The Sweep Amplifier ... 784
2.6 Circuits for Radar Range and Bearing Measurement ... 789
2.61 Electronic Bearing Marker Circuits ... 790
2.62 Range Marker Circuits ... 792
Fixed Range Markers ... 792

Variable Range Marker Circuits 793
2.7 Automatic Frequency Control and Automatic Gain Control 800
2.71 Automatic Frequency Control 800
Function and Requirements 802
AFC Circuit Design Considerations 804
Typical AFC Circuit Designs 805
2.72 Automatic Gain Control 809
2.8 Radar Receiver Power Supplies 810
2.81 Primary Power Sources 810
2.82 Low Voltage Power Supplies 811
2.83 High Voltage Power Supplies 814
Conclusion 816

Introduction

THE spectacular development of radar during World War II remains an outstanding achievement in the history of communications and the allied electronic sciences. With military necessity furnishing the required driving force and through the full interchange of technical knowledge among all interested workers in this field, it has been possible to extend our visual senses far beyond the horizons considered quite inelastic only a few years ago. The potentialities of radar in the peacetime world and the future application of radar design principles and techniques to the communications and allied fields justify a review of some further details of this wartime development.

The performance and design aspects of radar receivers will be considered in this paper. For this purpose, the radar receiver will be defined as that assemblage of components within the radar system which is required to detect, amplify, and present the desired information as gathered at the radar location. The input signals to the radar receiver consist of radio-frequency pulses containing information regarding the area under observation by the radar system, together with coordinate data defining further characteristics of this observed area. The output of the radar receiver is most commonly an optical presentation of this composite information, but in certain applications the output is further converted into electrical or mechanical signals for specific use. In general, the output of a radar receiver is presented in a form capable of immediate analysis and use.

Though the functional boundaries of the radar receiver are by the above definition quite distinct, the exact detailed composition of the receiver and the specific component designs are influenced by a considerable number of factors. The successful performance of a radar receiver is dependent to a large degree on the nicety with which these individual components are assembled into the system as a whole.

A study of the principal factors which influence the design of the radar receiver is presented in the following sections, followed by a more detailed exposition of the principal design aspects of the various components associated within the receiver. Illustrative equipment descriptions are included

where military security permits. All of the specific equipment examples presented here have been chosen from radar systems that have been developed within the Bell Telephone Laboratories and manufactured for the services by the Western Electric Company. This latter limitation excludes many interesting experimental developments which have not been produced in quantity and which have not, therefore, been substantially employed by the services during the war period.

It should be observed that the rapid development and successful employment of radar systems during World War II have come about through the cooperation and coordination of various governmental and military agencies, many research and development organizations, and countless individual workers. It is, therefore, an impossible task to assign individual credit for the details of the development which is here described. Radar has reached its present state of development through the efforts of many, not only those employed specifically on radar projects during the war years, but also those technical workers in the communications and allied fields in the years prior to the war who so adequately supplied the firm basic foundation upon which to build.

1. Radar Receiver Design Considerations

1.1 *The Military Radar System*

The specific use and area of operation of the radar system are two basic factors which exercise a profound influence on the receiver design. It is, therefore, pertinent to consider some common classifications of radar systems as employed during the war years.

A convenient functional classification of military radar systems may be made as follows:

A. *Search or Navigation*

This classification may include warning of the presence of enemy surface vessels or aircraft, navigation by location of landmarks, and reconnaissance.

B. *Missile Control*

This function includes radar systems to control gunfire and release bombs or missiles.

C. *Aircraft Interception*

This classification may be considered as an airborne combination of search and missile control, but is separated here because of the special radar design problems encountered.

As the detailed electrical performance requirements are primarily in-

fluenced by the above functional classification of radar systems, the mechanical equipment design or arrangement is likewise considerably influenced by the area of use of the radar system. As an illustration of this factor, radar systems may be alternatively classified according to the area of operation as follows:

A. *Ground Equipment*

The equipment design of a ground radar system must include provisions for portability, protection against damage in movement over rough terrain, and for operation under extreme weather conditions.

B. *Naval Surface Vessel Equipment*

Radar equipment employed on surface vessels is subject to extreme atmospheric corrosive conditions, severe shock from gunfire and, in some cases, partial immersion in sea water. Large naval vessel installations often involve extreme distances between the location of the various components of the radar system.

C. *Airborne Equipment*

Mechanical equipment features for aircraft use must include provision for operation over rapidly fluctuating conditions of temperature and atmospheric pressure. Vibration conditions coupled with strict weight requirements result in many additional problems from the equipment designer's standpoint.

Figures 1, 2, and 3 indicate typical radar equipments employed on the ground, sea, and air, some component designs of which will be reviewed in later sections of this paper.

1.2 *The Function of the Radar Receiver*

The basic function of a radar receiver is to translate and present the information received at the radar location in a desirable form. This requires that the receiver contain provisions to enable:

A. Conversion of the received signals, originally of a microwave frequency character, to signals in a frequency region more convenient to utilize.

B. Amplification of the extremely low energy signals as received to amplitudes useful to the observer.

C. Correlation of all other available pertinent data with the received microwave signals to allow determination of the complete coordinates and other desired characteristics of the target under observation.

Fig. 1.—Radar Equipment—Mark 20. This mobile equipment operating at 1000 mc is employed for searchlight control purposes.

D. Presentation of all desired information to the observer in a form capable of immediate analysis and use.

With these functional requirements in mind, it is in order to examine the character of the information available at the input and that required at the output terminals of a radar receiver.

1.21 *Characteristics of the Radar Receiver Input Signal*

In general, the signal radiated by the radar transmitting antenna and received by the receiving antenna consists of intermittent pulses of energy

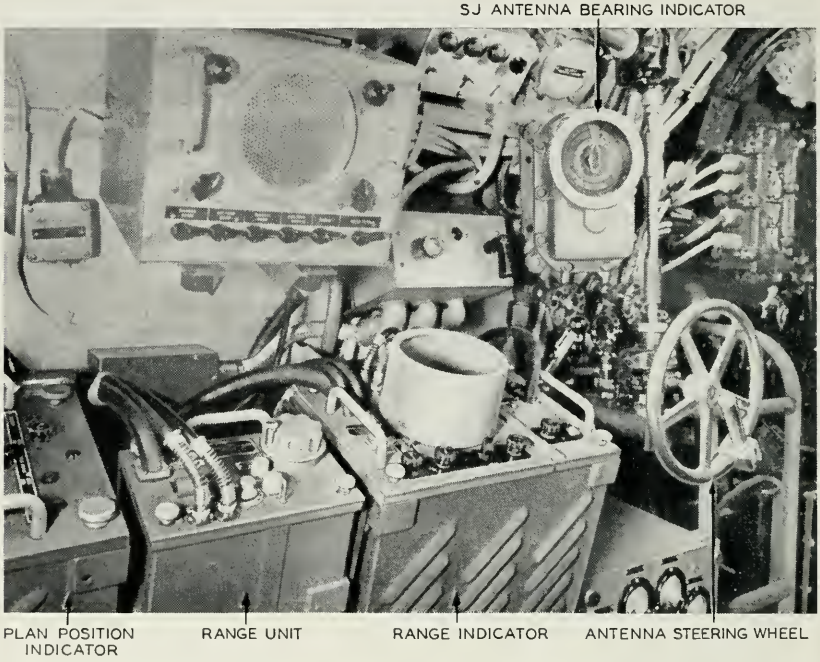


Fig. 2.—SJ Submarine Radar. Operating position of radar equipment in conning tower of U. S. Navy submarine.

at microwave frequencies. The methods employed in the generation and propagation of these radar microwave signals have been described elsewhere.[1, 2] For our purpose, it suffices to state that microwave frequencies extending from 700 mc to 10,000 mc are commonly employed. Pulse widths of 0.5 microsecond to 5 microseconds at repetition rates extending from 100 pps to 2000 pps are encountered in modern military radar systems, the

[1] "The Magnetron as a Generator of Centimeter Waves," J. B. Fisk, H. D. Hagstrum and P. L. Hartman, *Bell System Technical Journal*, Vol. XXV, April 1946.

[2] "Radar Antennas," H. T. Friis and W. D. Lewis, *Bell System Technical Journal*, Vol. XXVI, April 1947.

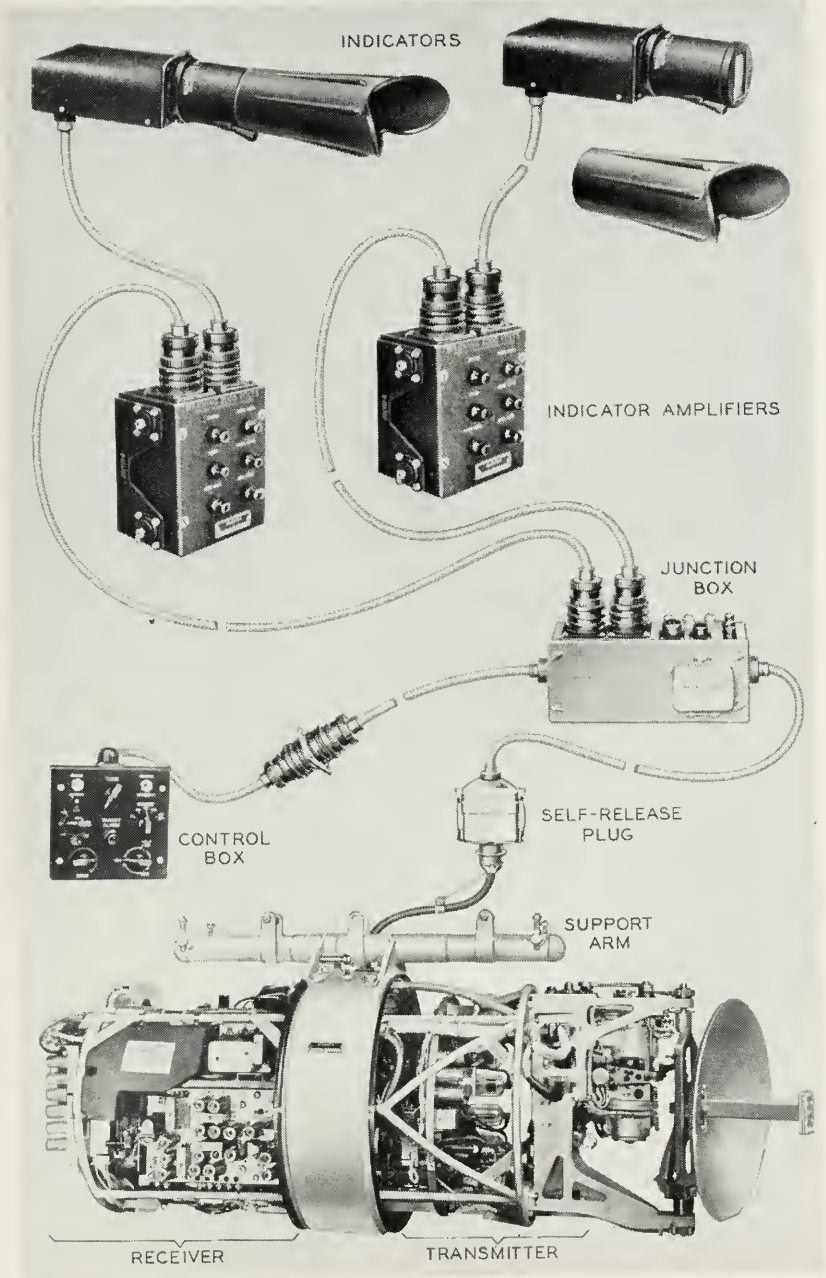


Fig. 3.—Components of Lightweight AN/APS-4. Airborne search and interception radar equipment.

exact choice of these parameters being dictated by the specific application of each type of equipment.

It has been customary to employ a common antenna system for both the transmitting and receiving functions, the necessary protection of the sensitive receiver input circuits from the high power transmitted pulse being furnished by a TR switch or tube.[3] The gas discharge TR switch assembly attenuates the energy fed to the input terminals of the receiver for the duration of the microwave outgoing pulse. At the time of decay of this transmitting pulse, the TR switch is arranged to offer low attenuation between the antenna and the receiver to any received signal.

The received microwave signal will be found to fluctuate in amplitude between extremely wide limits. This amplitude characteristic of a received radar signal is affected by the size and composition of the target, the power of the radiated outgoing pulse, the distance or range to the target, and miscellaneous propagation effects. Military requirements necessitate designing the radar receiver to perform successfully with the minimum received signal, while not unduly compromising this performance when signals of relatively high energy content are encountered.

Other characteristics of the transmitted microwave signal, such as pulse shape and repetition rate, are chosen to enable maximum performance to be attained for the specific application to be covered. The proper treatment of these miscellaneous characteristics of the radar signal is of basic concern to the receiver designer.

The primary basic information which can be derived from the characteristics of the received radar signal itself consists of data concerning the range to the target under observation. This range data is made available by a measurement of the elapsed time between the departure of the outgoing pulse and its return after reflection from the target and consideration of the velocity of electromagnetic wave propagation.

To determine the complete coordinates of a radar target, correlation of the range information, as determined above, and the direction of radiation from the antenna is necessary. Signals containing information as to the instantaneous attitude of the antenna with respect to chosen reference axes are, therefore, to be considered as essential inputs to the radar receiver.

Though the natural coordinate system of radar is of a polar form, many specific applications of radar systems require conversion of this information into other forms more convenient of use. For example, while in gun-pointing radar applications, it is desirable to present the final information in a polar coordinate form, corresponding to the aiming axes of the guns in many airborne radar bombing systems, it is necessary that the presentation be

[3] "The Gas-Discharge Transmit-Receive Switch," A. L. Samuel, J. W. Clark and W. W. Mumford, *Bell System Technical Journal*, Vol. XXV, January 1946.

made in terms of rectangular or other convenient coordinate systems referred to the ground itself. In certain applications, the characteristics of the display or presentation device requires that coordinate conversion functions be included within the radar receiver. The conversion and proper presentation of all radar system coordinate information will, therefore, be considered a necessary function of the radar receiver.

Additional forms of radar receiver input signals encountered are those primarily associated with the specific application of the radar system. Reference coordinate axes data obtained from compasses or gyroscopes are among the most common of these. In gun-fire control and bombing applications a considerable quantity of computed data must be accepted by the receiver. These data may include predicted quantities which must be presented in addition to the usual received present-time radar information. In the case of airborne radar systems, provision must be included to properly display navigational beacon and identification signals. The beacon is operated by the radar transmitter in the aircraft and returns a coded signal at a frequency slightly removed from the normal radar band. All aircraft radar receivers are required to adequately detect and properly display this information. It has also become common practice to require provision within the radar receiver for display of interrogator-response signals as employed for military identification purposes. The identification equipment (IFF) proper is not a radar system component and, therefore, it will not be considered here.

1.22 *Character of the Output of a Radar Receiver*

The output of a radar receiver is required to be available to the observer in a form which will permit immediate analysis and use of a maximum of the received information. The consideration of some additional characteristics of the radar information available and the military applications will furnish a basis for choice of presentation means in the receiver.

Because of the inherent pace of the constantly changing tactical military scene, the basic requirements imposed upon the radar presentation device are severe. For example, the use of radar in an aircraft, or on the ground or sea directed against aircraft involves a process of obtaining information on targets having relative velocities upward of 500 feet per second. If the radar system under consideration is being employed to furnish data for the release of bombs or to direct gunfire, a fraction of a second represents dimensions comparable to the target size. Such considerations adequately emphasize the extreme importance of retaining the "immediacy" characteristic of the information through the presentation device.

Another factor influencing the design of the presentation components in a radar receiver is found in consideration of the extreme complexity of the

received radar information. This complexity is created out of the comparatively limited resolution available and from the military need for presentation of detailed information concerning large areas during small time intervals.

The effect of limited resolution on the choice of presentation means of a radar receiver can be appreciated by considering the following. The microwave pulse employed in modern radar systems has a duration in time corresponding in linear range dimensions of hundreds of feet, while the beam width of commonly employed radar antenna systems likewise includes hundreds of feet of target at useful ranges. Thus, the inherent radio-frequency "resolution" is limiting to an extent that, while it usually enables one to determine the coordinates of the centroid of the target, it will not furnish adequate information as regards the exact size or shape characteristics of the target. The radar response of an area of military interest is a function of electrical conductivity and other related characteristics, rather than of the military importance of the target. These factors indicate strongly that the human observer must be required to supply a certain function of interpretation, and that the chosen radar presentation means should be such that this is possible. An effective illustration of this situation is found in the descriptive term "navigation by constellation" which was common among the radar operators in the long-range bombing forces. Here the navigation to and the orientation with respect to the military objective was often possible only through the interpretation of strong radar responses from known landmarks. Offset bombing, where the bombing radar operator carried out his observations on a satisfactory radar target in the vicinity of the final objective and introduced the known offset coordinates in the computed release point so as to strike the military objective, was found to be a successful method of partially overcoming this basic limitation of World War II radar equipments.

Modern warfare is concerned with rapid movement and extremely large area operations. The display of continuous information regarding these large areas is a basic military requirement of the modern radar system. For specific military applications the radar viewpoint may often be restricted to selected limited areas with increased demands on detail and on reproduction of changing information during small time intervals.

The above considerations have led to a choice of presentation means for the military radar system which is of an optical nature and has the essential characteristics of motion pictures. Such a display of complex information, in general, allows the observer to concentrate his attention at any time on any desired region of interest, to orient himself with respect to the broad features of the complete area, and to be cognizant of changes in the scene as they occur.

The cathode-ray tube has been most universally employed as the display device in the modern radar system. The incoming electrical information from the various receiver inputs is here electronically converted into a visual form capable of being modified over small intervals of time as the change in the radar scene occurs. Multiple presentations having various map-scale factors are found in many modern radar systems to enable detailed examination of a small magnified target area, while retaining the ability to observe the broad area features at will.

The use of radar for the purpose of control of gunfire, release of bombs, or steering of a vessel or aircraft requires that essentially continuous coordinate information be transmitted from the radar observer to the device under his control. This is usually accomplished through the registration of mechanical or projected electronic markers upon the visual radar display, this process of successive alignment furnishing the required information to the controlled device. In the case where automatic means of maintaining coincidence between marker and target are employed it should be observed that the original selection of the target and the initial coincidence adjustment still remains a matter for interpretation on the part of the human operator, and here again the visual radar display form is desirable.

The presentation means of a radar receiver has, therefore, been chosen to allow complete display of the data as quickly as it is received and in a form most convenient to the understanding of the human operator. In a broad sense, the radar system output terminal conditions and requirements are similar to those encountered for any communication system i.e. to supply the human observer with all the information available to the system in a form which will permit maximum usefulness with a minimum of delay.

1.3 *Composition of the Radar Receiver*

For convenience in the discussion to follow, the generalized radar receiver will be partitioned as shown in Figure 4. This particular choice of component division is somewhat arbitrary, but is chosen because of its functional simplicity. Mechanical, and in some cases electrical, conditions encountered in any particular radar system design often indicate physical arrangements which will differ considerably from the arrangement illustrated.

The converter component of the radar receiver has, as its primary function, the conversion of the received microwave signal to an intermediate frequency region where further amplification and discrimination is possible. The converter consists of a beating oscillator operating in the microwave frequency region and a nonlinear element, which at the higher radar frequencies consists of a point-contact crystal element. At the lower radar

frequencies it is customary to employ vacuum tube radio-frequency amplifiers preceding the converter element, and in these cases, similar vacuum tubes are employed as the nonlinear element. The output frequency of the converter commonly ranges from 30 mc to 100 mc. Because of the comparative difficulties experienced in the transmission of microwave energy over transmission lines or waveguides as contrasted with the problem at the lower intermediate frequency region, it is standard practice to locate the converter in close proximity to the antenna and transmitter portions of the radar system.

The intermediate frequency (IF) amplifier and associated second detector unit following the converter in Fig. 4 is required to obtain the necessary

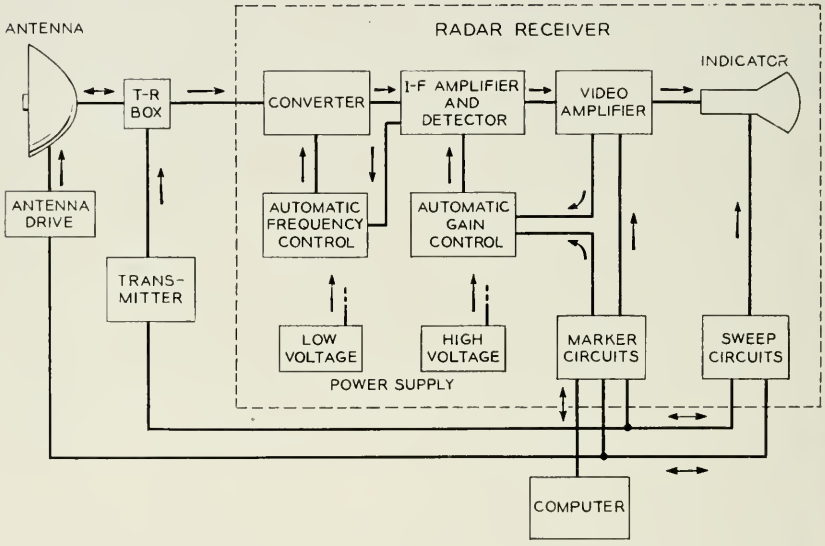


Fig. 4.—Schematic diagram of principal components of a military radar receiver.

amplification to the wanted signal, to supply discrimination against unwanted signals, and to finally convert the desired signal to a video form for presentation purposes. The usual gain required of a modern radar intermediate amplifier is of the order of 100 db. The band width of the IF amplifier is usually chosen between 1 mc and 10 mc depending on the specific radar system requirements. The techniques of construction developed for high gain radar IF amplifiers have resulted in compact component designs which are complete units in themselves, capable of being integrated into various radar systems.

The video amplifier characteristics are dependent to a large degree on the particular type of display system associated with it. Its primary func-

tion is to amplify the video signal output of the second detector, together with some other signals to be impressed on the cathode-ray tube. In this field the principles of design follow closely those developed for television prior to the war. Band widths of the order of 1 mc to 5 mc are commonly employed and output voltages ranging from 10 volts to 250 volts are required by the cathode-ray tube. Controllable nonlinear amplitude-gain characteristics are occasionally included in these video circuits to enhance the contrast or improve the apparent signal-to-noise performance.

The display device universally employed in modern military radar systems is the cathode-ray tube. The electro-optical response characteristics of this device may be chosen over quite wide limits to suit the specific needs of a particular system. In slow scanning systems, use of long time-decay optical characteristics of the sensitive screen is found to be advantageous, while in other cases of high-speed scanning, shorter persistence-type screens are employed. Screen diameters commonly employed range from 2″ to 12″ with a wide variety of color characteristics available to fit the detailed requirement of the radar system. A variety of radar presentation forms are employed to display most conveniently the received information for the specific application at hand. These display types differ primarily in the manner in which the radar field coordinates are presented. The deflection methods employed may be of an electrostatic or magnetic nature with combinations of each occasionally encountered.

The sweep circuit components of the radar receiver generate the electrical time wave forms necessary to display the received data properly. Here again the television art has supplied a technical background for these specialized electronic circuits. The great number of display types employed, requiring varied wave forms, has resulted in the development of a myriad of specialized sweep circuits whose apparent complexity is the result of varied combinations of elemental electronic circuits.

The range and time-marker circuits are required to interpret the coordinate data available to the receiver and to prepare this data for display in the desired form. Here again, television techniques have been employed and enlarged upon as the radar systems became more complex. A large number of specialized electronic circuit forms has resulted.

The automatic frequency control (AFC) and automatic gain control (AGC) components of a radar receiver have a function not unlike those elements found in most radio-communication systems. The automatic adjustments of the tuning and gain of the radar receiver has contributed greatly to the successful employment of radar under practical military conditions and have now become practically indispensable in the art. The circuit designs follow, in general, the techniques previously employed in radio communications though it will be observed that the character of the signal

and the close association of the receiver and the transmitter in the radar system introduces some additional new considerations.

The power supply components are shown divided into two types "low voltage" and "high voltage." This is a convenience which is desirable because of the different design problems encountered and the quite different equipment and circuits employed. The low d-c voltages required for the operation of the electronic components of a radar system vary from 100 volts to 500 volts with both polarities with respect to ground often required. The cathode-ray tube and "keep alive" circuits of the TR tube require voltages from 1000 to 7000 volts usually at low current. The voltage regulation of these power supply components to permit stable radar system operation under the extreme and variable military operating conditions encountered presents a problem of considerable magnitude to the radar receiver designer.

2. Radar Receiver Component Design

2.1 *The Radar Receiver Input Circuit*

The conversion of the received microwave radar signal to a lower frequency region where more efficient amplification is possible represents an extremely important function of the radar receiver. The basic military requirement of radar, that of providing the greatest possible useful sensitivity, depends fundamentally on the efficient handling of the low-energy microwave received signal in the input of the radar receiver. The microwave character of the received signal and the extremely low amplitudes encountered contribute to the difficulties of radar input circuit design.

The techniques of microwave transmission available to the communications engineer have but recently been developed and are at this time still extremely limited in comparison to those methods commonly employed at the normal radio frequencies. Even short physical connections required between elements in the microwave region become "electrically long," a matter of several wavelengths in the usual case, and here the design problem becomes acute. Network element of inductance, capcitance, and resistance are not available in the form so efficiently employed at the lower frequencies. Waveguides and cavities at radar frequencies replace them, and the design of selective frequency networks in the microwave region becomes a matter of precise arrangement and control of complex mechanical forms.

Suitable means for vacuum tube amplification at frequencies above 1000 mc have not been available to date; this represents another extremely restricting situation for the radar receiver designer.

2.11 *Input Signal Characteristics*

The low amplitude of the signal at the input terminals of the radar receiver requires that this signal be efficiently utilized. The power of the received signal at this point under somewhat idealized free space assumptions is given by the following:

$$P_R = \frac{G^2 P A_\epsilon^2}{16\pi^2 D^4}$$

Where G = Power gain of common transmitting and receiving antenna
P = Transmitted power of radar system
$A\epsilon$ = Equivalent flat plate area of target (This represents an equivalent flat plate normal to the incident beam which reradiates all impinging energy)
D = Range to the target.

The two following sample computations are illustrative of the military radar system conditions:

1. Naval Vessel Search Radar System
   Frequency = 3000 mc
   Target range = 25 nautical miles
   Target-Destroyer (Effective flat plate area = 0.03 sq meter at 3000 mc)
   (This value has been determined from a study of target response with military radar systems)
   Power gain of Antenna = 30 db
   Transmitter Peak Power = 100 kw
   Received Peak Power = 13×10^{-14} watts
2. Airborne Search Radar System
   Frequency = 10,000 mc
   Target Range = 70 nautical miles
   Target-Destroyer (Effective flat plate area = 0.2 sq meters at 10,000 mc)
   Power Gain of Antenna = 30 db
   Transmitter Peak Power = 100 kw
   Received Peak Power = 11×10^{-14} watts

A reduction of the available received signal power, as computed above, is to be expected in practice due to multiple path effects and absorption and refraction effects over the propagation path.

2.12 *Input Circuit Noise Considerations*

While it is possible to conceive of providing sufficient gain within a radar receiver to meet any desired sensitivity requirement, this sensitivity cannot usefully be employed beyond certain limits as determined by the amplitude

of the noise disturbances at the input terminals of the receiver. Noise disturbances may be defined as the resultant unwanted interfering electrical energy at the point under consideration and includes contributions due to atmospheric disturbances, unwanted radiation from adjacent electrical equipment, microphonic disturbances, and noise due to vacuum tubes and thermal agitation. At microwave frequencies we are usually concerned only with the thermal agitation and tube noise contribution. Atmospheric disturbances at radar frequencies are negligible and microphonic and electrical interferences from adjacent electrical equipment can, by proper and sufficient engineering, be reduced to any desired level.

It has been shown[4, 5] that the thermal noise (Johnson noise) voltage which appears at the input terminals of a radio or radar receiver is determined by the value of the resistance component of the generator impedance at this point. For maximum transfer of signal power the load termination is required to be equal to the internal impedance of the generator and for this condition the total thermal noise power delivered to the load is given by

$$P_N = KTB \text{ (watts)}$$

where:

K = Boltzman's constant = 1.38×10^{-23} Joule/degree abs.

T = Absolute temperature in degrees

B = Bandwidth under consideration in cycles per second; and the signal-to-noise ratio is given by

$$\frac{P_S}{P_N} = \frac{P_S}{KTB} \text{ (a numeric)}$$

where P_S = maximum available signal power.

If the signal generator referred to is followed by any 4-terminal network representative of a converter element, an amplifier, or a passive network, the effective signal-to-noise ratio at the output terminals of the network will be modified. To obtain a measure of this effect we may assign a figure of merit, F, to the network called the "noise figure" of the network and define this as the ratio of the available signal-to-noise ratio at the signal generator terminals to the available signal-to-noise ratio at the output terminal of the network.

This may be written as:

$$F = \frac{P_S}{P_N} \frac{P_{S_2}}{P_{N_2}} = \frac{P_{N_2}}{KTBG} \text{ (a numeric)}$$

[4] "The Absolute Sensitivity of Radio Receivers," D. O. North, *R. C. A. Review*, Vol. VI, January 1942.

[5] "Noise Figures of Radio Receivers," H. T. Friis, *Proc. I. R. E.*, Vol. 32, No. 7, July 1944.

where:

$$G = \frac{P_{S_2}}{P_S}, \text{ by definition the "gain" of the network.}$$

or we may write:

$$P_{N_2} = FKTBG \text{ watts.}$$

It is common practice to express F in decibels given by the relation $10 \log_{10} F$.

For the case of two generalized networks in tandem following the signal generator and having the same effective bandwidth we may similarly write:

$$P_{n_{out}} = F_a G_a G_b KTB + (F_b - 1) G_b KTB$$

$$= \left(F_a + \frac{F_b - 1}{G_a} \right) G_a G_b KTB \text{ watts.}$$

where the subscripts refer to networks a and b.

The effective noise figure of such a system is given by:

$$F_{\text{system}} = \left(F_a + \frac{F_b - 1}{G_a} \right)$$

This expression indicates the importance of the gain of the first network in the over-all system noise performance.

As an illustration, a noise figure of 11 db and a loss of 6 db may be considered as acceptable performance for a typical silicon crystal converter operating at 3000 mc. If the following input circuit of the IF amplifier has an effective noise performance represented by a noise figure of 5db , the over-all system noise figure will be found to be 13 db. The reduction in system performance due to the noise contribution of the input stage of the IF amplifier here is approximately 2 db. If the system performance must be improved by increasing the power of radiation, the importance of this secondary noise contribution is apparent.

A comparison of the noise figures of a point-contact crystal converter element and the GL-2C40 Lighthouse vacuum tube used as an amplifier and as a converter element is given in Fig. 5. At frequencies below 1000 mc there is a definite advantage in employing a vacuum tube as a radio-frequency (RF) amplifier preceding the nonlinear element.

2.13 *1000* MC *Radio-Frequency Amplifier Design*

In the design of military radar systems for the 1000 mc operating range the GL-2C40 vacuum tube has been employed rather universally in converter circuits.[6] The essential design features of this special purpose triode

[6] "The Lighthouse Tube," E. D. McArthur and E. F. Peterson, *Proc. of National Electronics Conference*, Vol. 1, 1944.

are illustrated in Fig. 6. The basic advantage of the GL-2C40 tube for use at high frequencies is found in its construction, whereby the tube elements are arranged to form an integral portion of the external circuit with a minimum of mechanical disturbance. At these frequencies external coupling circuits of the transmission line type are usually employed. This tube when operated at an anode potential of 250 volts has a mutual conductance of approximately 6000 micromhos and an amplification factor of 35. It has been customary to employ one or two stages of RF amplification associated

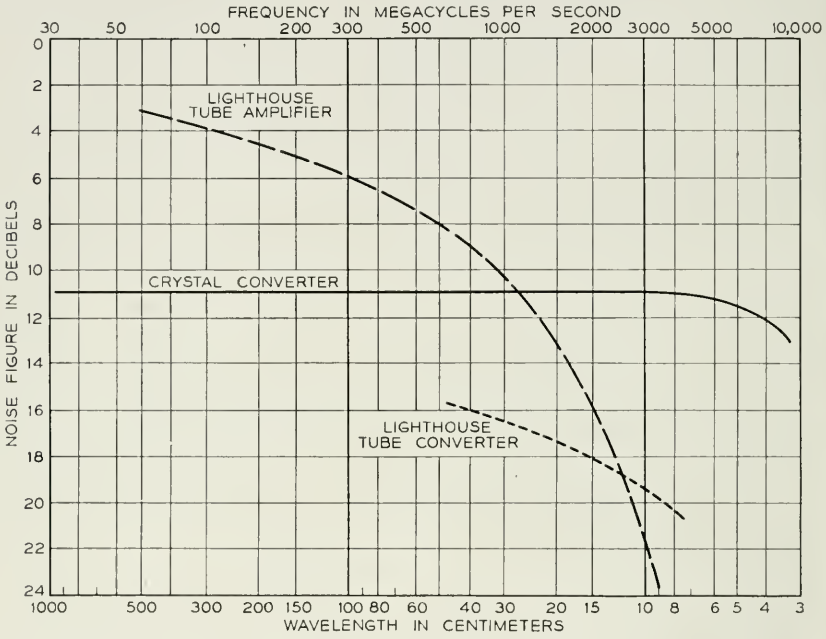


Fig. 5.—Comparison of noise figures for point-contact silicon crystal rectifier and GL2C40 vacuum tube.

with the nonlinear element and a beating oscillator using this same tube. The reduced performance of the GL-2C40 tube as a converter element as indicated in Fig. 5 is not of importance, if sufficient gain is provided prior to the actual conversion process, and the ability of this vacuum tube to operate at higher levels is a positive advantage.

The electrical circuit design techniques employed in this frequency region are based on the use of transmission line elements in place of the more or less orthodox lumped element configurations at the lower frequencies. The difficulties experienced in practical designs of radar converters of this type revolve about successfully terminating and satisfactorily isolating each stage so that confusing and inefficient interaction effects are minimized.

The simplified schematic of a typical radio-frequency amplifier which employs a GL-2C40 vacuum tube operating at approximately 1000 mc is given in Fig. 7, together with an outline of the mechanical arrangement of the tuning elements. This amplifier is of the tuned-grid, tuned-plate type

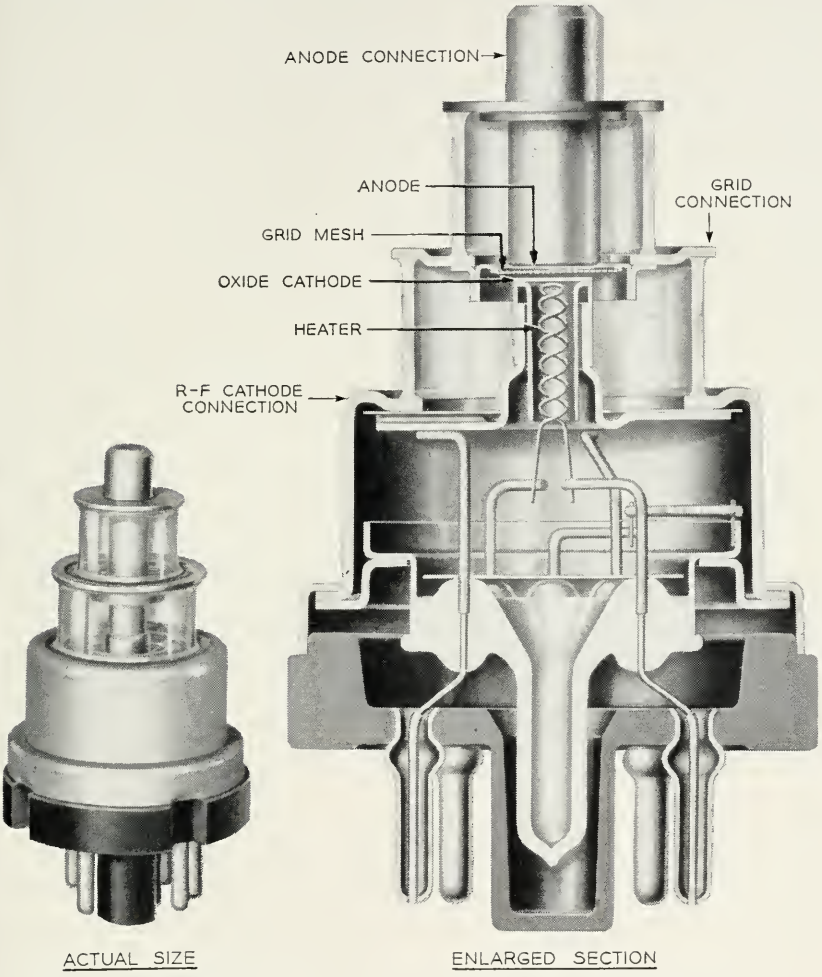


Fig. 6.—Constructional Features of the GL-2C40 (Lighthouse) vacuum tube.

of circuit where the input and output circuits consist of coaxial transmission line elements tuned by sliding plunger or ring elements, and is so constructed to include the GL-2C40 tube as an integral part of the tuning structure. The input coupling to the amplifier is achieved by means of a probe extend-

ing into the input or grid cavity and is proportioned for operating out of a 50-ohm impedance cable connection to the radar antenna system. The output coupling is obtained from a coupling loop located in the region between the grid and plate concentric sleeves. The gain of this RF amplifier design is 12 db when operating at a frequency of approximately 1000 mc and the noise figure of this component is 14 db.

2.14 *The Radar Converter*

The basic converter is illustrated in Fig. 8 and may be defined as a device which has two input pairs and one output pair of terminals and is charac-

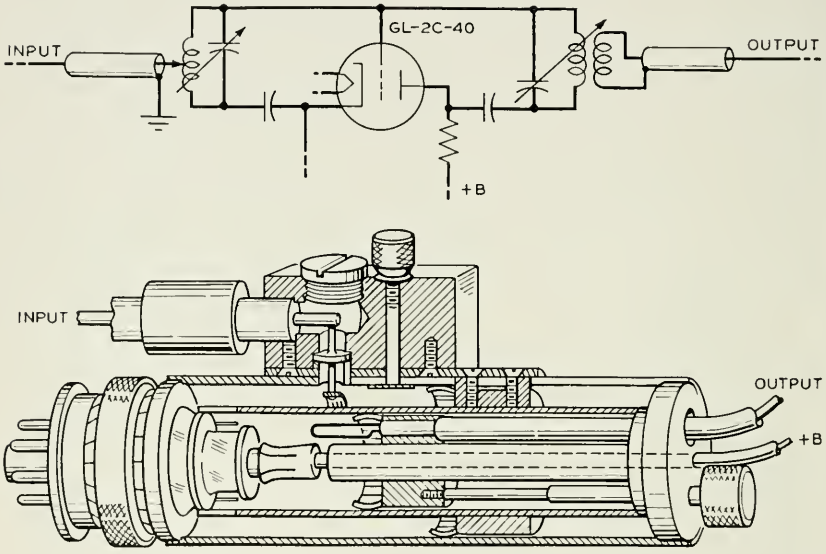


Fig. 7.—1000-mc Radio Frequency Amplifier. Simplified schematic diagram.

terized further by the property of delivering an output signal, which, in terms of amplitude of one of the input signals, is essentially linear but which has an output frequency related to the sum or difference of the two input frequencies. This frequency conversion is obtained by the use of an element which has a nonlinear voltage-current relationship and upon which is impressed the two input signals. The desired sum or difference frequency signal is then selected and utilized as the wanted signal output.

The basic configuration of the converter shown in Fig. 8 employs a nonlinear element with network coupling means to provide efficient transfer of signal power into and out of the component and a beating oscillator with its associated coupling network to supply the required additional input fre-

quency. One or more of the networks involved may, of course, be arranged alternatively in a parallel configuration if desired. The nonlinear element may be a vacuum tube with properly chosen operating conditions or a point-contact silicon crystal rectifier. The beating oscillator in the modern military radar converter employs special types of vacuum tubes such as the GL-2C40 type previously mentioned or the single-cavity velocity modulated types[7] at the higher radar frequencies.

A typical voltage-current characteristic of a point-contact silicon rectifier is shown in Fig. 9. This nonlinear characteristic may be expressed mathematically as a power series with coefficients whose amplitude decreases quite rapidly with the order of the associated term. If two sinusoidal voltages represented by frequency f_1 and f_2 are simultaneously impressed

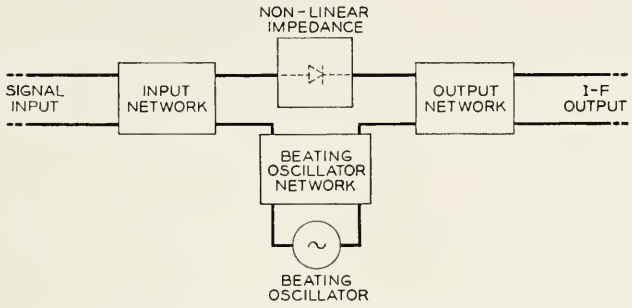


Fig. 8.—Basic configuration of a radar converter.

upon such a nonlinear element, the resultant current flowing in the output impedance will produce output signals of the form $nf_1 \pm mf_2$ where n and m are integers including zero. The effective amplitude of each of these modulation products is related to the magnitude of the power series coefficients. In the radar converter under consideration f_1 may represent the received signal frequency and f_2 represents the beat oscillator frequency. The difference terms of the above expression are of the greatest importance here because the desired output signal frequency has been chosen at a low value as compared with the input received radar signal. The selection of the wanted first-order difference term is accomplished by the frequency selectivity characteristics of the converter output coupling network and the following IF amplifier.

A major design problem encountered in the practical development of microwave radar converters is one concerning the design of the coupling networks. The previously referred to limitations of microwave network

[7] "Reflex Oscillators," J. R. Pierce and W. G. Shepherd, *Bell System Technical Journal*, Vol. XXVI, July 1947.

techniques and the difficulty in maintaining precise control of each element over the required band of frequencies are the fundamental problems of the radar converter designer. The network problem here is concerned with realizing the desired frequency conversion with a minimum of dissipation of the useful signal. This implies that coupling of the nonlinear element to the input and output terminals must be achieved in such a fashion that matched impedance conditions result for the signals in their respective frequency regions. Since the output signal has been shown to appear as a

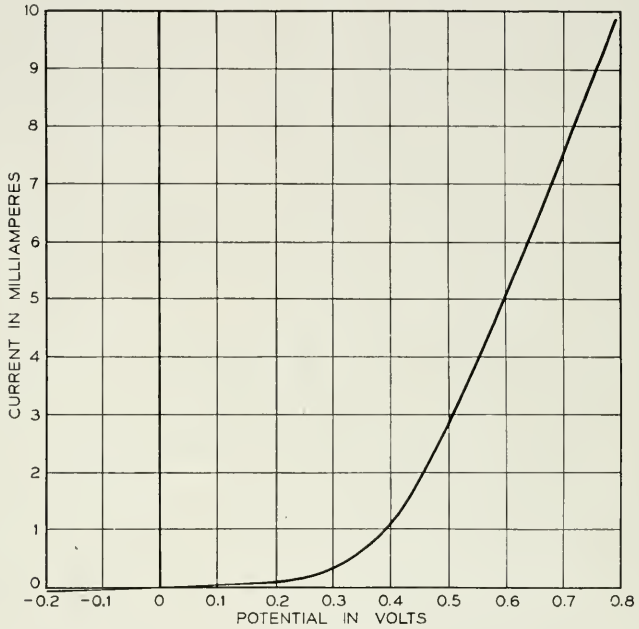


Fig. 9.—Typical voltage-current characteristic of a point-contact silicon crystal rectifier.

number of energy concentrations extending over a wide frequency range, it would appear that an efficient design of output coupling network would involve optimizing the impedance relationships over this wide-frequency band. Several factors tend to simplify this problem by restricting the output frequency band which must be considered in a practical radar converter design. First, the greatest amount of energy is found to be present in the lower order modulation products because of the rapid reduction in amplitude of the higher order terms in the equivalent expression for the nonlinear element. This results in a concentration of energy at the input, output, and beat oscillator frequencies and their respective second harmonic regions. Second, the ratio of f_1 to f_2 in a microwave radar converter is of necessity

essentially equal to one, this factor also contributing to a narrowing of the frequency regions of interest to those around the input, the beat oscillator, and the output values. The third factor, which is of assistance to the converter designer, is the effective separation of the input and output circuits by the loss of the nonlinear element. Where a vacuum tube is employed as the nonlinear element, the interaction of these circuits may be made quite negligible and, in the case of the crystal rectifier the inherent loss of this element, which may be of the order of 6 db and undesirable from a radar system performance standpoint, does simplify the converter network design.

It has been found in practice that it is sufficient to consider the impedance conditions at the internal terminals of the converter networks in the frequency regions which include $f_1, f_1 - f_2, f_1 + f_2$ and $2f_2 - f_1$.

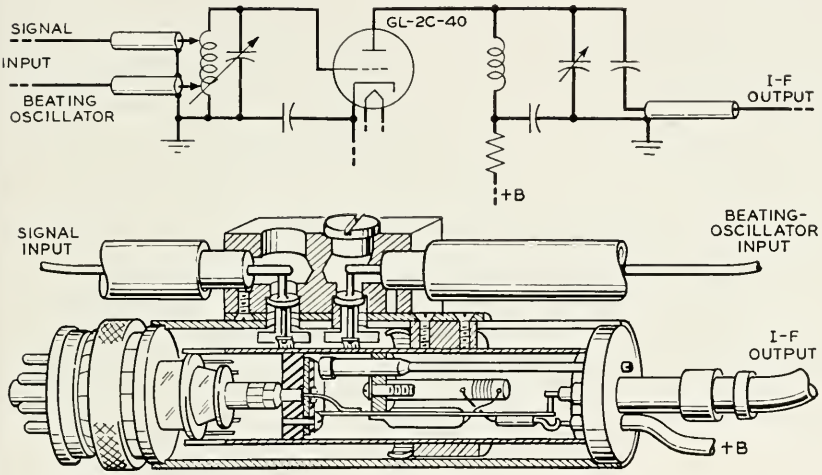


Fig. 10.—1000-mc Vacuum Tube Radar Converter. Simplified schematic diagram.

The matter of efficient transfer of power from the local beating oscillator to the nonlinear element is of secondary importance generally because of the relatively large amount of power available. This condition is advantageous allowing mismatch loss in this branch to effectively minimize the unwanted interaction effects with input and output circuits.

Figure 10 illustrates the schematic and certain constructional features of a vacuum-tube converter which has an operating frequency of approximately 1000 mc. The similarity of circuit and mechanical arrangement to that of the radio-frequency amplifier unit, shown in Fig. 7, is apparent. In the case of the GL-2C40 converter the two input frequencies are similarly probe coupled to the grid-cathode circuit, maintaining optimum impedance conditions for the signal input and locating the beat oscillator probe so as

to effect an impedance mismatch and thus reduce the interaction between this circuit and the input circuit. The output coupling network, in this case operating at 60 mc, consists of an inductive impedance tuned with a variable condenser. The output of this converter is fed into the following IF amplifier by means of a 75-ohm coaxial transmission line. The loss of this converter unit, defined as the ratio of the power of the wanted 60-mc output signal to the signal power impressed upon the input terminal, is

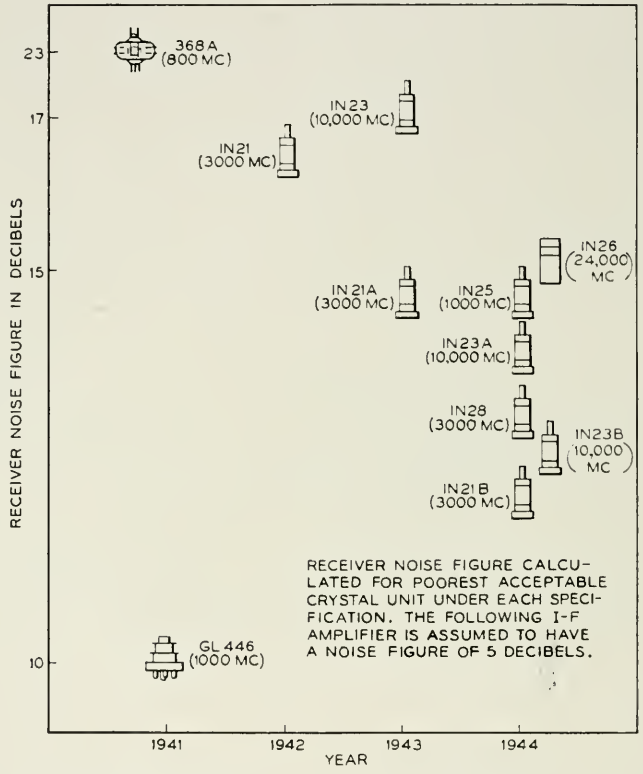


Fig. 11.—Chronological development of point-contact silicon crystal rectifiers and associated receiver noise performance.

approximately 6 db. The noise figure of a typical unit of this type is approximately 20 db, this value being of secondary importance since sufficient gain is provided in the associated radio-frequency amplifier.

At the higher radar frequencies the noise performance characteristics of silicon crystal-point contact rectifiers are considerably better than that of vacuum tubes available during the past war years. Figure 11 outlines the chronological developments of these units with respect to the receiver noise

performance. The performance of two types of vacuum tubes are included for comparison purposes. The development details of the silicon crystal unit have been discussed elsewhere.[8] This development, together with the corresponding magnetron and reflex oscillator studies, assumes a status of major importance in the progress of the development of military radar equipment during World War II, allowing the designers to extend the frequency of operation upward with increased system performance resulting.

It is sufficient to limit our attention to the frequency region 3000 mc and above in the consideration of the silicon crystal converter. It is apparent that the absence of suitable vacuum tube radio-frequency amplifiers in this region imposes a strict requirement on the efficiency of conversion of this component.

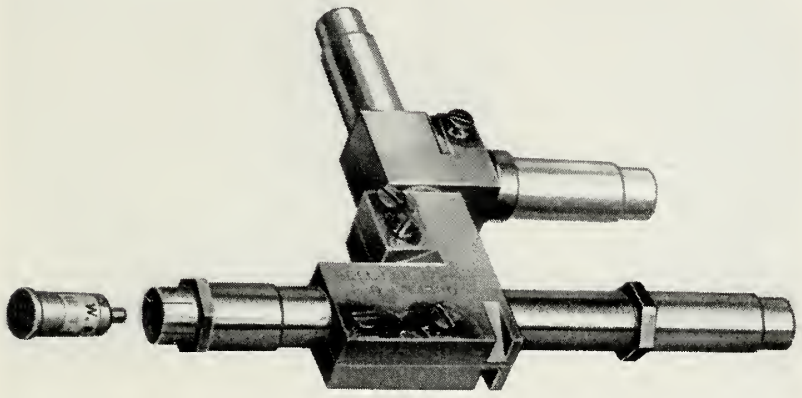


Fig. 12.—3000-mc Crystal Converter of an early design.

Historically, one of the first of the microwave converters of the silicon crystal type, operating at 3000 mc and which was employed in military radar equipment, is shown in Fig. 12 together with an illustration of the mechanical arrangement[9] in Fig. 13. In this early model the silicon crystal element was mounted in a cartridge type unit, a method which proved quite satisfactory and was followed for the remainder of the war years. The input tuning circuit of the converter here illustrated consists of a coaxial transmission line element having a length of approximately three-quarters of a wavelength and adjustable to enable fine control of tuning. The nonlinear ele-

[8] "Development of Silicon Crystal Rectifiers for Microwave Radar Receivers," J. H. Scaff and R. S. Ohl, *Bell System Technical Journal*, Vol. XXVI, January 1947.

[9] "Microwave Converters," C. F. Edwards, To be published in a forthcoming issue of *Proceedings I.R.E.*

ment is located at the high-impedance end of this transmission line, while the other end is essentially short-circuited at the input frequency by means of a small by-pass condenser element built into the IF output transmission line. The input line is coupled into this tuned circuit by means of a variable coupling probe and the local beat oscillator in turn is similarly coupled to the input line. The point of coupling of the beat frequency oscillator input is so arranged as to introduce an effective mismatch and thus provide adequate isolation of this and the input circuit. The output circuit of this converter includes the by-pass condenser previously referred to, together with the input transformer of the first IF amplifier stage. The average loss of a

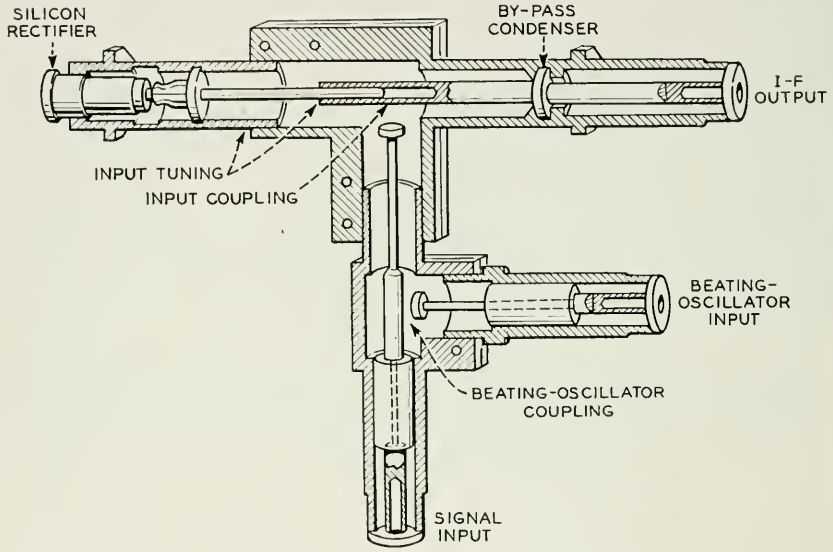


Fig. 13.—3000-mc Crystal Converter. Schematic diagram.

3000-mc crystal converter of this type was 6 db and a noise figure of 11 db was realized.

Another design of crystal converter which was developed during the early part of the war, and whose basic form was employed in many military radar equipments operating in the region of 10,000 mc, is shown in Fig. 14. Here the silicon crystal cartridge is positioned within the waveguide with its axis parallel to the E vector plane and at a point approximately one-quarter of a wavelength from the short circuiting piston which terminates this assembly. The IF output is obtained from the coaxial line mounting structure shown which offers a low impedance to the input frequency by virtue of its equivalence to a one-half wavelength element with a short circuit at the far end.

The dielectric supporting rings shown form an RF by-pass element minimizing the loss of input signal power in the IF output network. In this design the beating oscillator energy is introduced into the waveguide by mounting the reflex oscillator tubes adjacent to the waveguide in such a fashion that the output probe is inserted into the waveguide cavity at a point removed by an odd one-quarter of a wavelength from the face of the TR output iris. This assures reflection of the local oscillator energy which travels toward the TR tube directing it toward the crystal element. The degree of coupling of the reflex oscillator circuit is varied by adjustment of the distance that the probe is inserted within the waveguide. For airborne applications an additional oscillator tube is included for beacon reception. This basic form

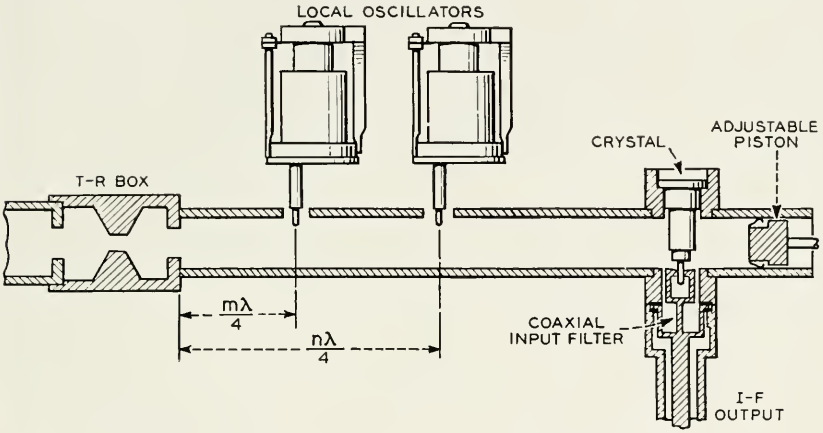


Fig. 14.—10,000-mc Crystal Converter. Schematic diagram.

of radar converter was employed in large numbers in the military airborne radar field during World War II.

A third type of crystal converter design which was developed in the latter period of the war is illustrated in Fig. 15. A basic difference in this structure is found in the use of a waveguide hybrid junction often referred to as a "magic tee." This junction has an electrical performance characteristic at microwaves similar to that of the hybrid coil common to low frequency communication circuits, i.e., a 4-pair terminal network with an internal configuration such that power applied to any one pair of terminals will appear equally at two other pairs of terminals, but will not be available at the remaining pair of terminals. Referring to Fig. 15, it should be noted that power applied to the input waveguide will appear equally in the output branches but is balanced out of the beat oscillator branch. In a similar fashion, the beat oscillator power will appear equally in the two output

branches but will not appear in the input waveguide. Certain impedance matching adjustments are obtained through the use of the matching rods positioned as shown.

The method of insertion of the crystal cartridge into the waveguide in this

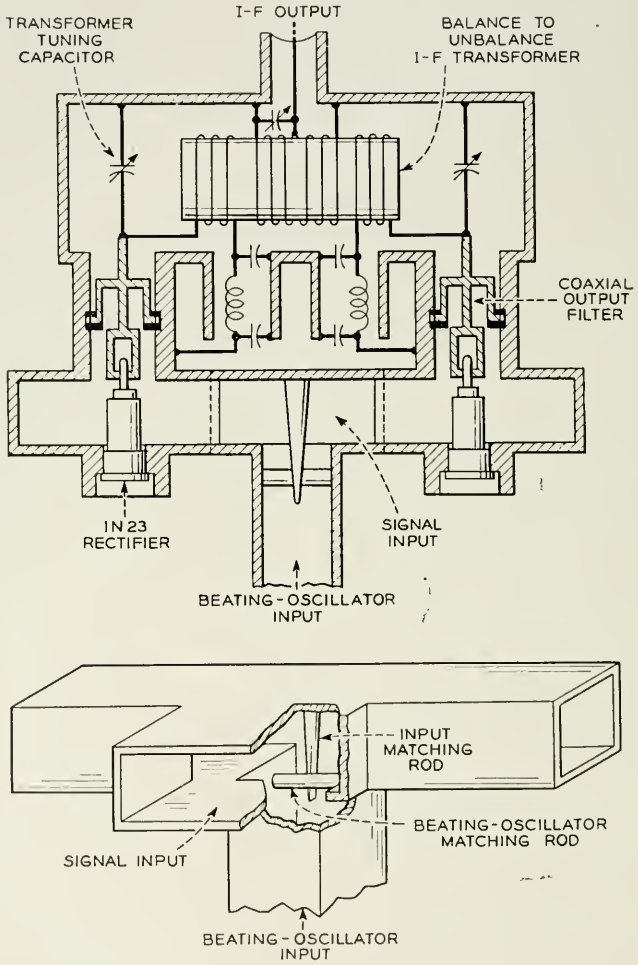


Fig. 15.—Balanced crystal converter employing wave-guide hybrid junction.

design follows the scheme employed in the converter just described. However, here two crystal elements are employed in a balanced form which necessitates a balanced-to-unbalanced impedance transformation of the IF output signal for transmission over a coaxial line to the input stage of the IF amplifier. The degree of balance obtainable here is necessarily a func-

tion of the similarity of the crystal elements as well as the other elements shown. In practice, no particular difficulty was experienced in maintenance of sufficient balance with the improved production control of crystals during the latter part of the war program.

The advantages of the balanced radar converter here described are twofold. First, the signal power dissipation in the beating oscillator branch is reduced to a minimum and conversely the beating oscillator power fed back into the antenna and reradiated is reduced. Second, the noise sidebands, which are associated with the output frequency of the reflex oscillator, are reduced effectively in the IF output branch by the degree of balance available. This local oscillator noise sideband contribution is normally responsible for a definite degradation of the over-all radar receiver noise performance and hence, the use of a balanced converter will contribute to improved performance. An additional advantage of the balanced converter is the minimizing of signal branch impedance variation effects on the beating oscillator load impedance and, therefore, its frequency. The variation of the antenna impedance during the scanning cycle has in this design little effect on the tuning of the receiver.

2.15 *The Radar Receiver Beat Oscillator*

For microwave radar receiver purposes the selection of the local beat oscillator within the converter assembly has been essentially limited to two types of tubes, both of which were developed during the past war period. For radar systems operating at frequencies of 1000 mc and lower, the GL-2C40 lighthouse triode has served quite satisfactorily, while at frequencies above 1000 mc, the single-cavity reflex oscillator tube has been extensively employed. Both of these tube types have adequate power output and frequency stability characteristics to meet the normal radar system requirements.

Some of the desirable characteristics of a beat oscillator tube for use in military radar receivers can be listed as follows:

A. At least 20 milliwatts of useful output power is desirable. In the case of the silicon crystal rectifier element, the applied power is limited to approximately 1 mw; however, the availability of additional oscillator power enables the converter designer to effectively isolate the beat oscillator and signal branches by simple inpedance mismatch means.

B. The frequency stability of the ideal beat oscillator tube must be inherently good, or convenient means to automatically control this frequency must be provided. The maximum allowable radar receiver frequency variation due to all causes is of the order of 1 mc. In terms of the beating oscillator frequencies employed in the radar systems of the past war, this represents an allowable oscillator frequency variation

from all causes of from 0.1% to .01%. The possible influencing factors include temperature, atmospheric pressure, supply voltage, and load impedance variations with time, and mechanical shock and vibration.

C. It is extremely desirable that frequency control of the beat oscillator be available by remote electrical means. The use of automatic tuning control of a military radar receiver has proved a necessity during the past war and, as will be discussed in a later section, the rates of change of frequency encountered are found to be quite great. This requires essentially that an electrical control method of continuously adjusting the beat oscillator frequency be employed to obtain satisfactory receiver performance.

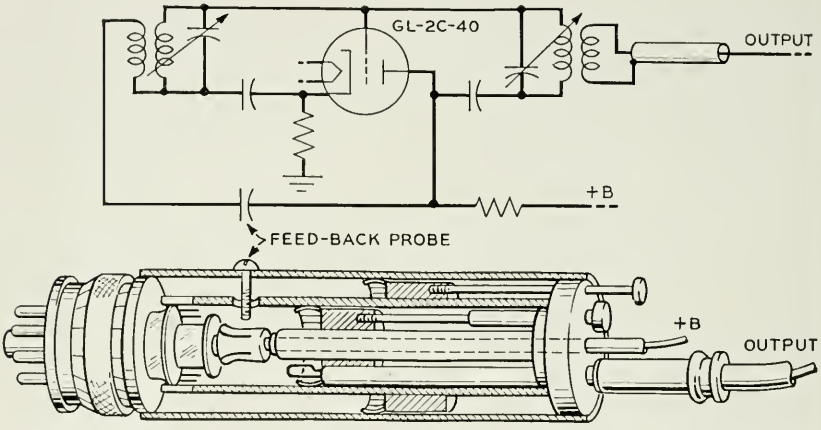


Fig. 16.—1000-mc Radar local beat oscillator employing GL-2C40 vacuum tube—schematic diagram.

D. It is desirable that the output of the beating oscillator tube be free from noise. In the usual radar system, if the output frequency of the beat oscillator is modulated with noise, a reduction in receiver performance will result. As previously discussed, the development of the balanced converter has provided the converter designer with some relief from this noise source and in this case this requirement assumes less importance.

A beat oscillator arrangement utilizing the GL-2C40 lighthouse tube, as developed for a military radar system operating in the 1000 mc region, is indicated in Figure 16. This assembly is quite similar mechanically to the RF and converter components employed in this frequency region and described previously. The positive feedback necessary to sustain oscillation is provided by means of a feedback coupling probe as shown. The oscillator output is available by means of a pick-up loop inserted into the

plate-grid cavity and thence to the output coaxial line. The frequency stability of this particular design has been quite satisfactory due to the considerable development effort expended on the mechanical design features of this assembly.

Another radar receiver beating oscillator which has been extensively employed in military equipment designs operating at 3000 mc and higher during the past war period is the reflex velocity-modulated oscillator. The 2K25 type, which is a typical single-cavity reflex oscillator for operation in the 10,000-mc region, is illustrated in Fig. 17. The general principle of operation is quite straightforward, but the complete theory of operation is exceedingly complex and has been described in detail elsewhere.[10] A beam of electrons of relatively uniform velocity and density is projected from a cathode surface toward a cavity space defined in part by two grids and then toward a repeller electrode beyond. The presence of the oscillatory potential between the two cavity grids acts to impart initial velocities to the electron stream in accordance with the cavity radio-frequency potentials existing at the time of crossing of this gap. Under certain operating conditions between these grids and the repeller electrode, which is maintained at a negative potential, "bunching" of electrons occur and upon the return of the electrons to the cavity region under the retarding influence of the repeller electrode, a certain amount of power may be extracted and utilized externally. As might be expected from this cycle of events, the optimum operating conditions necessary for reinforcement of oscillation within the cavity are related to the time required to return the electrons to the cavity with reference to the instantaneous oscillatory radio-frequency potential of the cavity. Thus, numerous modes of oscillation are found in this type of reflex velocity-modulated oscillator which are related to each other by integral numbers of periods of the oscillatory frequency and the transit time of the electrons. In the practical application of the reflex oscillator the number of useful modes are limited to perhaps two or three, the external power output available at the additional theoretical modes being reduced by dissipative conditions within the oscillatory system. The relation of repeller potential to the appearance of these modes is illustrated in Fig. 18 for the case of a 2K25-type oscillator. It should be observed here that the frequency at the maximum power output condition for each mode is the same, i.e. the frequency associated with the cavity dimensions for that series of modes. The cavity dimensions of the reflex oscillator tube are varied by the application of external mechanical pressure regulated by an adjustable tuning strut as shown in Fig. 17. The power for external use is obtained from a coupling loop located within the cavity region and transmitted through the base by means of a coaxial lead.

[10] Pierce and Shepherd, Loc. Cit.

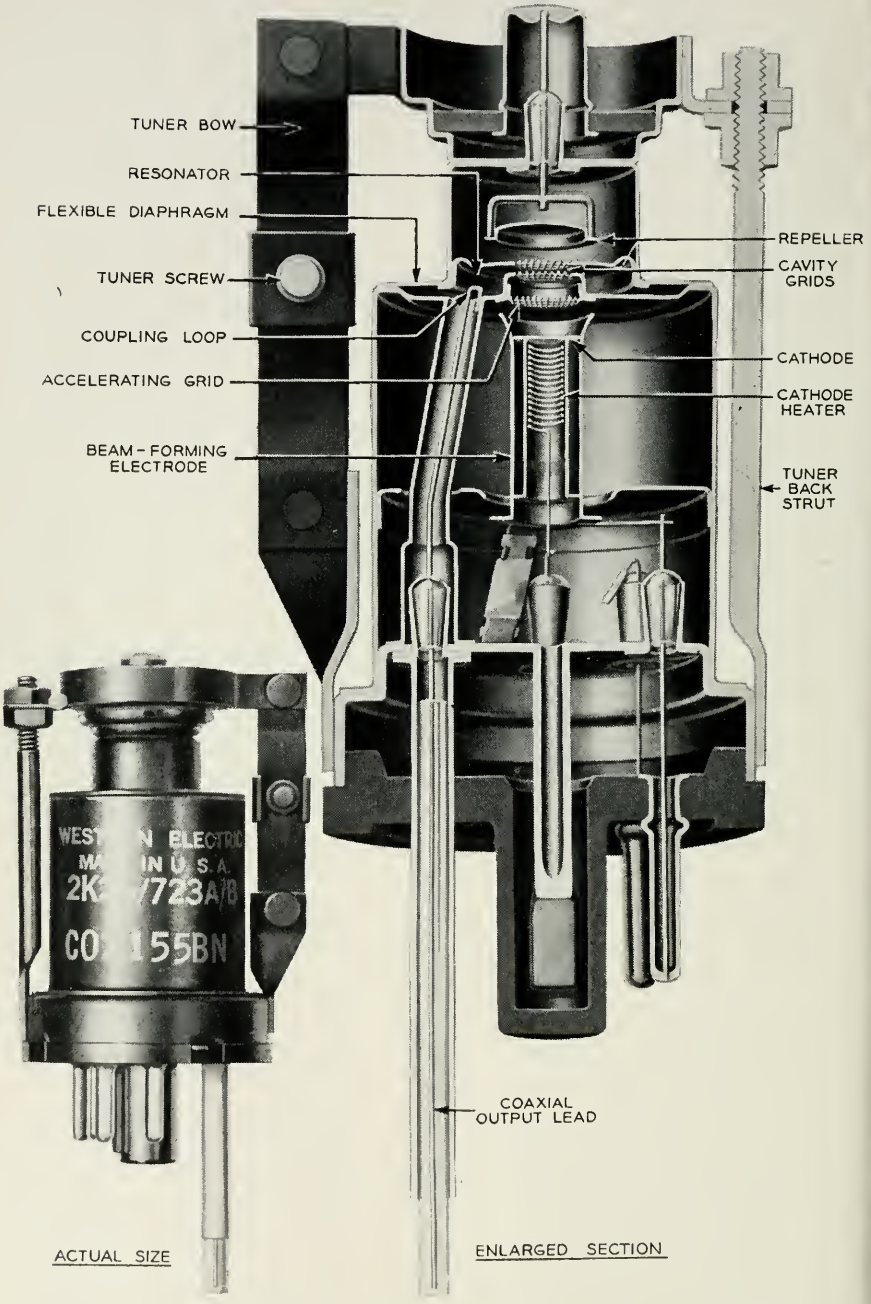


Fig. 17.—Constructional details of the type 2K25 single-cavity reflex oscillator for use at 10,000 mc.

The single-cavity velocity-modulated oscillator is admirably suited to electrical remote control of its oscillation frequency by means of the potential applied to the repeller element, thus lending itself to automatic frequency control in a simple manner. Figure 19 indicates the frequency and power output versus repeller voltage characteristic of a typical 2K25 10,000-mc reflex oscillator when operating at a normal mode previously shown in Fig. 18. It is customary to define the electronic tuning range of a reflex oscillator as that range of frequencies over which the power output exceeds

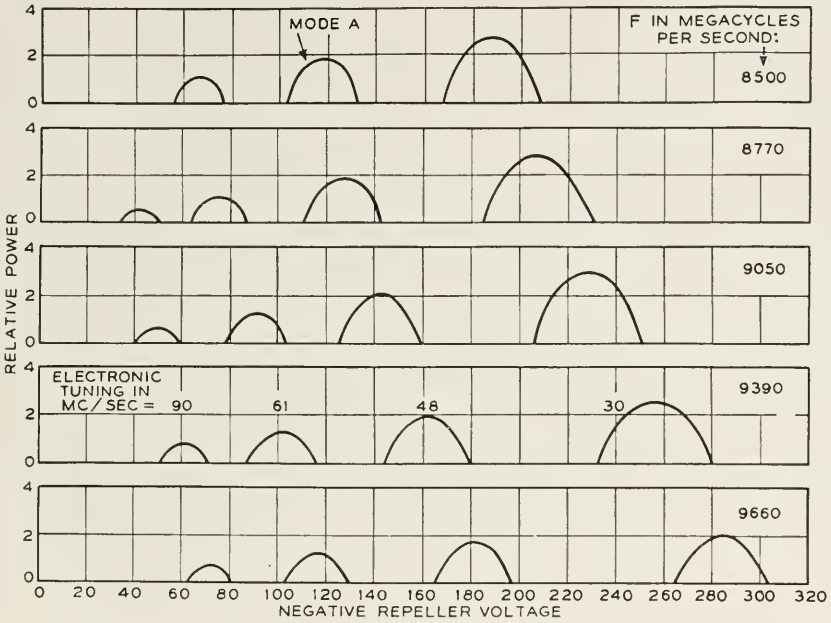


Fig. 18.—Typical output power modes vs. repeller voltage for a 2K25 type reflex oscillator.

50% of the maximum power output. The tube whose characteristic is illustrated in Fig. 19 would accordingly have an electronic tuning range of 53 mc.

2.16 *Typical Radar Input Circuit Designs*

An example of a radar receiver input circuit operating in the 1000-mc frequency range is shown in its final mechanical form in Fig. 20. This particular design was employed in several military ground radar systems including the Mark-20 Searchlight Control equipment of Fig. 1, and in modified form in several naval fire control systems. It consists of two stages of radio-frequency amplification, a converter stage, and a local beating

oscillator all of which employ the GL-2C40 lighthouse tube. This particular equipment example has been developed in the form of a sliding drawer to assure ease of maintenance but with no sacrifice in rigidity of mechanical assembly so necessary in this type of equipment. Further mechanical details of the cavity and tube structures of each component are shown in Fig. 21 and are generally similar to the examples previously described. These cavity structures are heavily silver-plated to assure good electrical and thermal conductivity. The problem of adequate conduction of heat from the GL-2C40 vacuum tube under the severe ambient temperatures

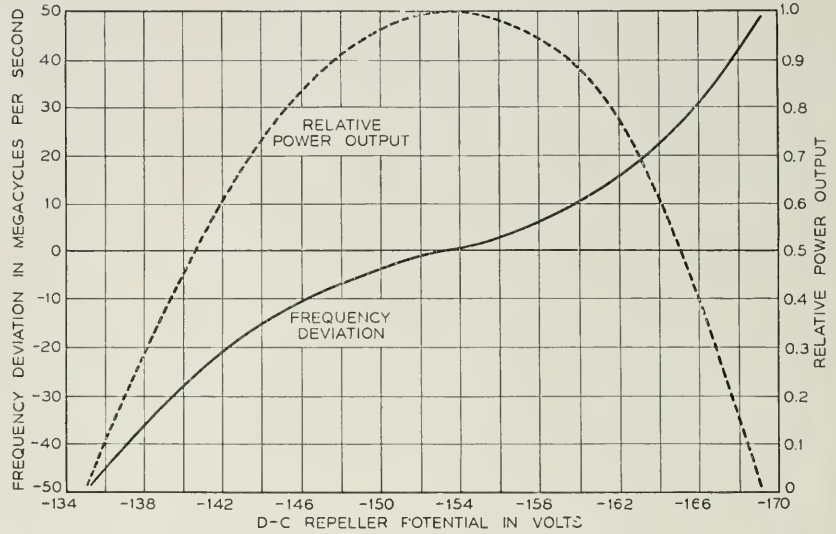


Fig. 19.—Power output and frequency deviation vs. repeller voltage for 2K25-type reflex oscillator.

encountered in military service is extremely important. Positive locking devices are provided for all adjustments in this mechanical design.

Figure 22 indicates the schematic diagram of this design. It may be observed that separation filters are employed extensively to assure negligible interaction effects between the various stages through the common power supply connections.

The over-all performance of this particular design of a radar receiver input circuit is as follows:

Input Frequency	1000 mc
Output Intermediate Frequency	60 mc
Over-all Band Width	5 mc
Over-all Gain	18 db
Over-all Noise Figure	14 db
Input Impedance	50 ohms
Output IF Impedance	75 ohms

The above performance is representative of the limits which all manufactured units are required to meet and also indicates the field performance which must be maintained for satisfactory military service. Under laboratory conditions and with a certain compromise of stability and ease of

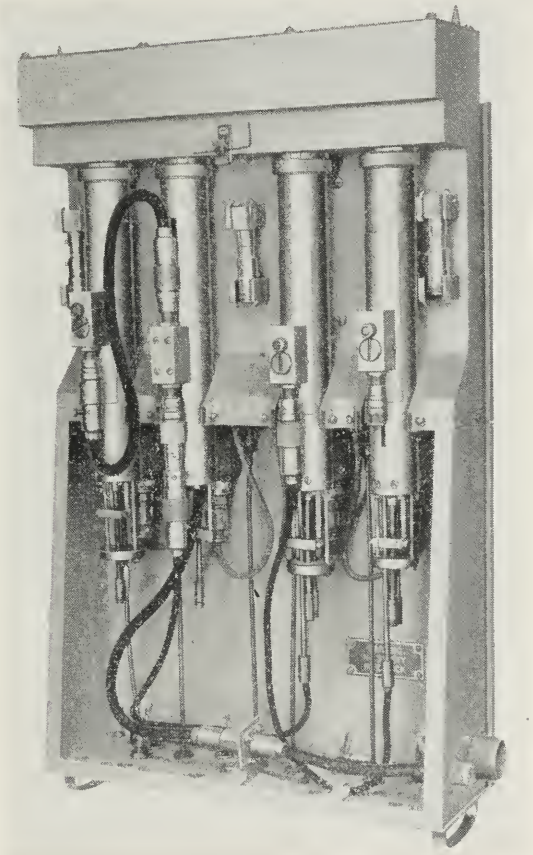


Fig. 20.—1000-mc Radar receiver input circuit design, including two stages of radio frequency amplification, converter, and beat oscillator.

adjustment, improved performance over the figures given here can be expected.

The design for manufacture of a converter assembly typical of airborne equipment methods is illustrated in Fig. 23. Here the equipment design reflects the fundamental requirement of radar equipment for aircraft—that of providing compact units of sufficient rigidity with a minimum of weight. This particular converter unit was employed in the AN/APS-4 Airborne

Search and Interception radar equipment shown in Fig. 3 and operates within the 10,000-mc frequency band. The automatic frequency control

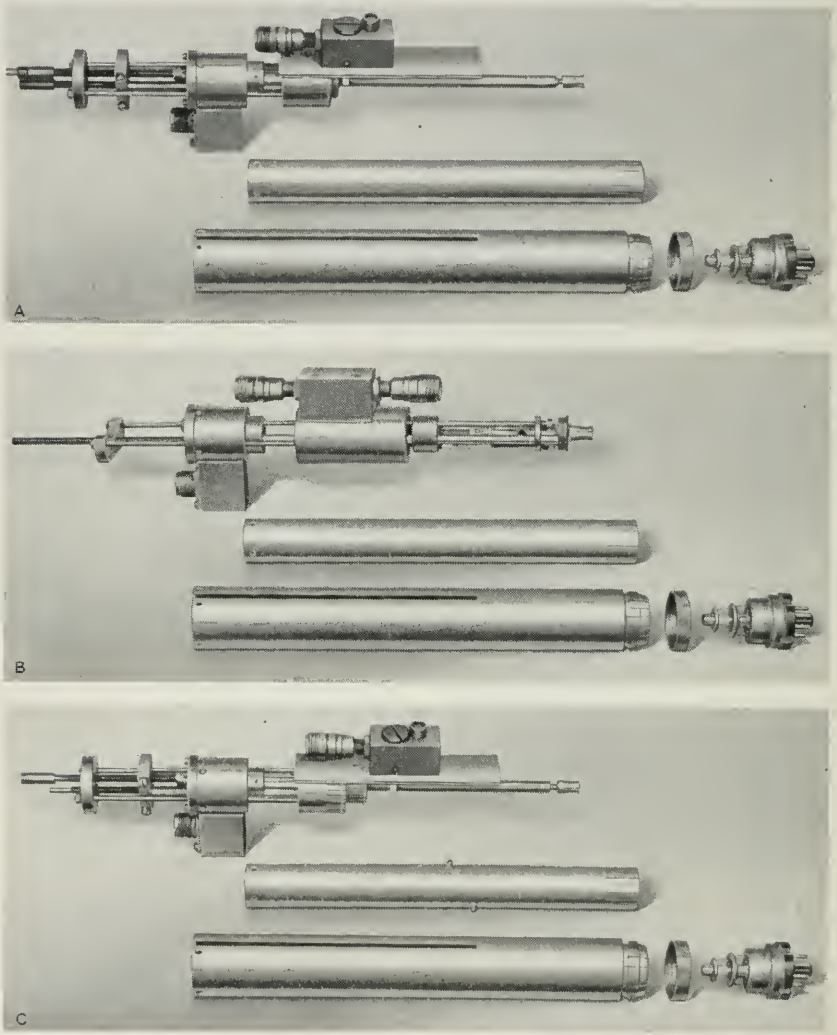


Fig. 21.—Mechanical details of cavity structures employed in 1000 mc radar receiver input circuit.

circuit and two *IF* preamplifier stages are also included here as an integral part of the converter assembly.

The schematic diagram of the converter portion of this assembly is given in Fig. 24. It should be observed that the basic arrangement of the crystal

converter is similar to a form described in a previous section. The signal in this example is introduced into the converter section of the waveguide through an iris following the TR tube. Two 2K25 reflex oscillator tubes are mounted upon this waveguide with their output probes extending into the guide proper. The crystal cartridge is located near the end of the waveguide with an adjustable piston terminating the guide. A waveguide to

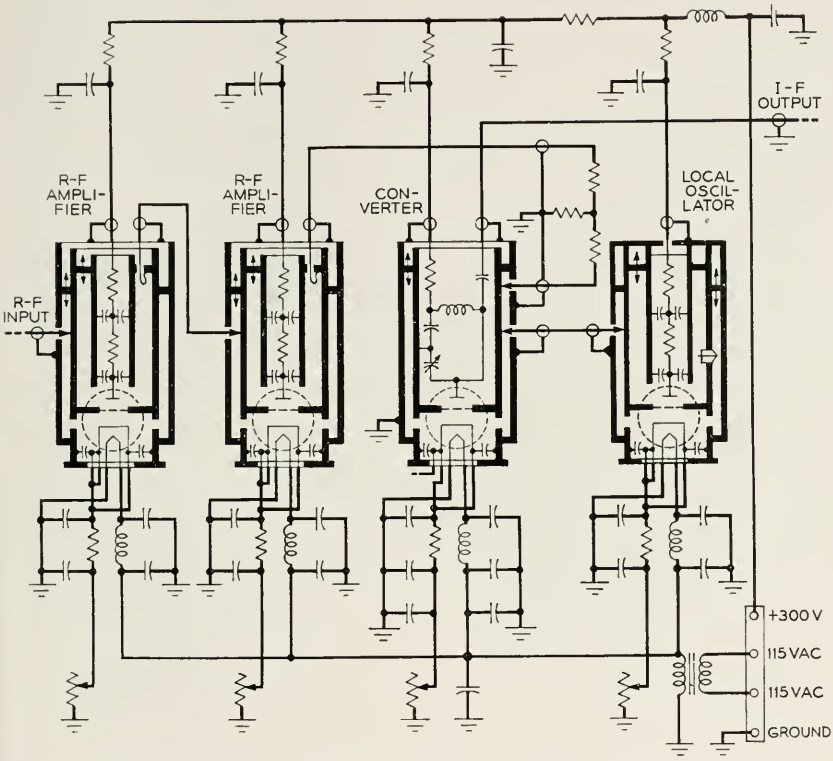


Fig. 22.—1000-mc Radar Receiver Input Circuit. Schematic diagram.

coaxial transforming section employs the crystal as an extension of its axial element with an adjustable capacitance element for tuning purposes.

It will be observed here that a shutter is included in this design, whereby the crystal element can be isolated from the signal input end of the waveguide. This is a most necessary device on all radar converters employing crystal converter elements to prevent accidental overload of the crystal. When the transmitter is not in operation and accordingly the TR tube "keep alive" is not energized, there is a possibility of subjecting the crystal element to signals of sufficient amplitude to destroy its characteristic. These over-

load signals may occur as the result of direct pick-up of an adjacent radar system or in certain cases atmospheric discharges. To prevent this a mechanical shutter is inserted into the converter which offers effective attenuation to any signal input. This shutter is withdrawn only if the proper TR "keep alive" voltage is available.

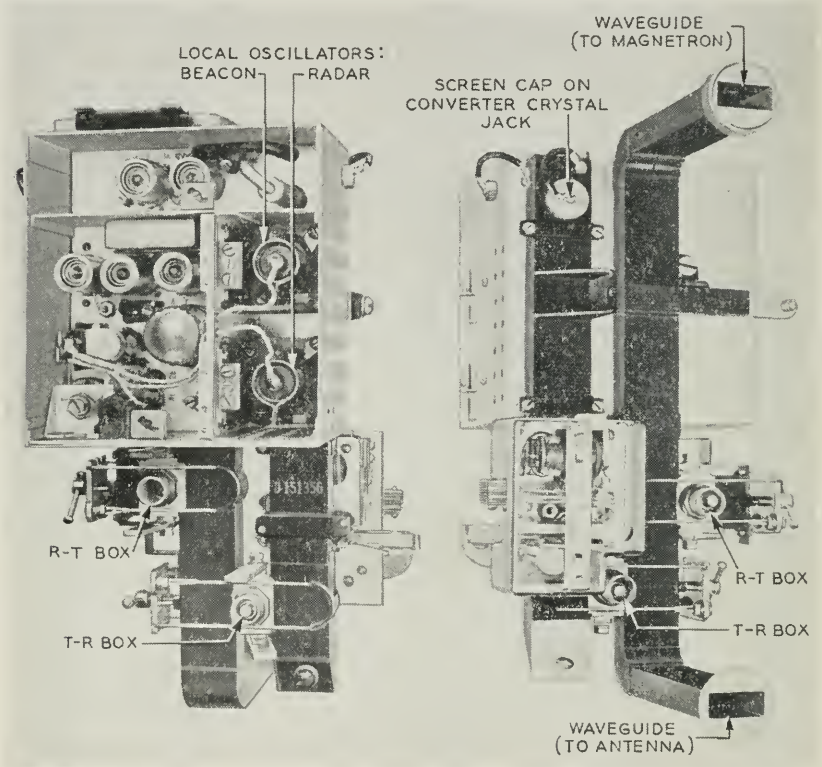


Fig. 23.—10,000-mc Converter and IF preamplifier assembly for AN/APS-4 airborne radar equipment.

The performance characteristics of this converter and IF preamplifier unit is as follows:

Radar Frequency	10,000 mc
IF Frequency	60 mc
Conversion Loss	6.5 db
Noise Figure of IF Preamplifier	4.5 db
Noise Figure of Converter	8.6 db
Noise Figure of Converter and Preamplifier	11.9 db
IF Band Width	4 mc

These values are representative of the performance under conditions of average crystals and average IF input tubes.

2.2 *The Radar Intermediate Frequency Amplifier*

The intermediate frequency (IF) amplifier component of the radar receiver has as its function the selection and amplification of the received signal following its conversion to the intermediate frequency. The further conversion of the IF signal to a video form suitable for use in the display device is usually included as an integral part of the IF amplifier and will, therefore, be considered here as an additional function of this component.

2.21 IF *Amplifier Requirements—Band Width*

The frequency-selectivity characteristic of the radar receiver is determined effectively by the IF amplifier, since the preceding converter and other microwave components offer little or no selectivity. The receiver

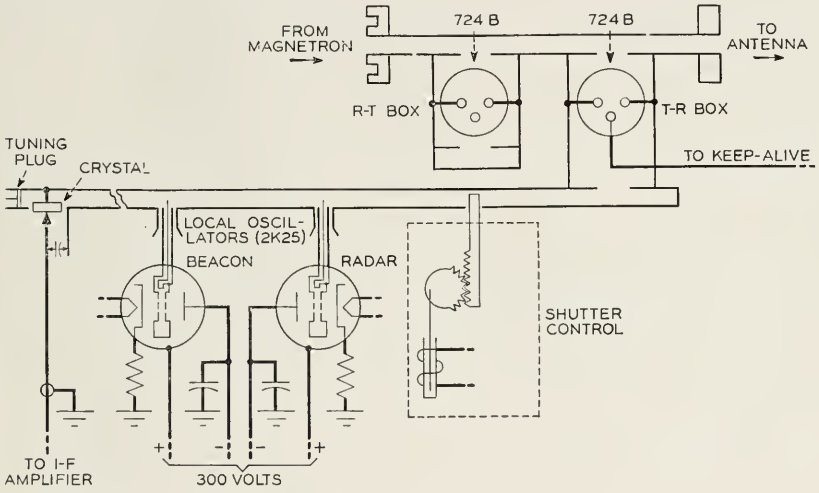


Fig. 24.—Schematic diagram of 10,000-mc AN/APS-4 converter unit.

band width required to adequately transmit the wanted signal, while restricting the noise contributions, is an extremely important factor in the radar system design and represents, therefore, a basic consideration in the design of the IF amplifier.

The receiver band width required to adequately transmit the basic received pulse information can be determined by a consideration of the frequency-energy characteristics of a radar pulse. The microwave pulse is created by the sudden application of a high-energy pulse to the magnetron microwave generator. It has been found sufficient to treat this output envelope as a simple rectangular pulse in band width computations. Figure 25 indicates the amplitude, time, and frequency relationships which exist for an idealized rectangular radar pulse envelope. It may be observed

that approximately 75% of the energy contained in the idealized original pulse will be available after transmission through a band-pass structure of band width dimension of $\frac{1}{\tau}$ cycles per second. A further doubling of the band width to $\frac{2}{\tau}$ cycles per second will increase the available energy by only about 15%. In the case of the practical radar-pulse envelope which usually is characterized by a trapezoidal form with finite rise and decay intervals, the energy contained at frequencies outside of a $\frac{1}{\tau}$ band is reduced somewhat over the idealized case illustrated in Fig. 25.

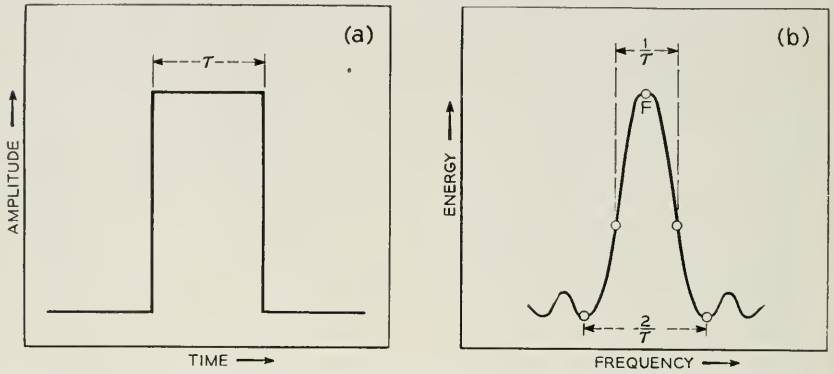


Fig. 25.—Amplitude-time and energy-frequency relationships for a rectangular radar pulse envelope.

The signal-to-noise ratio of the radar receiver is dependent on the over-all receiver band width as indicated in the previous section of this paper. It is extremely important then to restrict the IF band width as much as possible, consistent with adequate transmission of the signal itself. The final band width choice is that value where a further increase would result in a noise increase greater than the corresponding signal improvement and where a band width reduction would diminish the signal by a greater increment than the noise. The exact determination of the optimum radar receiver band width must be carried out using the final display device in making the signal-to-noise comparisons. If the radar system is to be employed for search purposes where echo presence is the primary measure of the performance of the equipment, the optimum over-all receiver band width has been found to be of the order of $\frac{1}{\tau}$ cycles per second resulting in an IF amplifier band width of $\frac{2}{\tau}$ cycles per second to adequately transmit the double-

sideband signal at this point. In the case of fire-control radar equipments, where precision range measurement is required, it is desirable to determine the range by reference to the leading or lagging edge of the radar received pulse. Here the optimum band width may be considerably greater than the value indicated above for a simple search system to assure a minimum rise time of the displayed pulse and an accordingly more precise determination of the position of the pulse. Additional factors which influence the radar receiver band width value in a particular system design involve the frequency stability of the microwave generator and the local beating oscillator, the sensitivity characteristics of the automatic frequency control system, the frequency stability characteristics of the IF amplifier itself, and finally the desire to permit ease of tuning by the radar operator.

The phase distortion introduced by the IF amplifier is of secondary interest in the case of radar systems. The faithful reproduction of all characteristics of the received radar pulse is usually not of extreme importance, since with few exceptions the criterion of presence is all important. The detailed form of the transmission characteristic is likewise not extremely critical, the usual "rounded" IF transmission characteristic, however, contributing somewhat to ease in tuning the radar receiver.

Gain Characteristics

A consideration of the converted input signal levels encountered and the video output level desired indicates the IF amplifier gain requirement. The input signal to the IF amplifier is determined by the type of converter employed and the presence or not of radio-frequency amplification preceding the IF section and the absolute noise power resulting. The video level at the second detector must be maintained at a sufficiently high level so that microphonic disturbances within the remaining video components are negligible and low enough to assure satisfactory detection without serious overload effects. Undesired feedback at the IF frequency also tends to limit the practical gain which can be introduced into the IF amplifier. In the military radar systems of the past war period the usual maximum gain associated with the IF amplifier was of the order of 110 db with a maximum detector output level of approximately 1 volt rms of input circuit noise. The extreme variation of the level of the desired radar signal makes necessary that provision for a large gain control variation be included in the IF amplifier design. This gain control often involves automatic, as well as manual, adjustment and commonly a gain variation range of the order of 80 db is required.

Another consideration which enters into the IF amplifier design and is associated with gain features, is the behavior of the amplifier in the presence of extremely large radar signals or enemy "jamming" signals. Optimum

protection against complete "blocking" of the IF amplifier under these conditions involves the use of extremely small-valued filter or by-pass elements which are associated with the grid and cathode circuits to present very short time constants and, accordingly, assure recovery of the amplifier in fractions of a microsecond after overloading by a large pulse signal. The gain control method is also chosen to minimize possible overloading by reducing the gain of the amplifier at a point as far forward in the radar receiver chain as possible.

Intermediate Midband Frequency

The choice of the intermediate frequency for a radar receiver is essentially a compromise between the need for reduction of unwanted external interference and noise, and the desire to realize maximum performance in terms of gain and noise figure.

The tendency to employ a high IF midband value arises from a consideration of the character of the certain local beating oscillator noise sidebands. As mentioned previously, the noise sideband output decreases as the frequency interval from the oscillator frequency is increased and it is apparent that a high value for the IF will be advantageous. In the case of balanced converters where the oscillator noise sidebands are reduced by the circuit balance, this factor assumes less importance. A moderately high IF is also advantageous in the elimination of the IF signal from the final detected video output which must be accomplished by the use of low-pass filters. The automatic frequency-control problem is somewhat simplified by the use of a high IF. The wider separation of the desired tuning point and the image response and the reduction of interfering TR pulse energy is a positive help in the performance of the automatic frequency-control circuits and will be discussed in detail in a later section.

There are also a number of factors which indicate that a low value of midband IF may be desirable. The noise figure of the input stage of the IF amplifier is generally better at a low frequency, though this improvement tends to be quite small for the band widths employed in the military radar system. A more important advantage of a low IF value is found in the improvement of the absolute frequency stability of the IF transmission characteristic under the influence of variations of tube and circuit capacitance.

Intermediate midband frequencies of 30 mc and 60 mc have been employed in the majority of military radar systems developed in the United States during World War II. In general, 30 mc has been employed quite extensively in naval and ground radar equipments for fire control where radar pulse widths of the order of one microsecond are used. For airborne radar equipments with the emphasis on compactness, weight, and the trend toward higher microwave transmission frequencies, 60 mc has proven to be an extremely popular intermediate frequency value.

The Second Detector

The final conversion of the IF signal to a video form is accomplished by simple detection. This process is usually associated with the IF amplifier because of the relative ease by which the video signal can be transmitted to the following display circuits, usually physically removed from the IF amplifier, as compared with the transmission problem which exists at the intermediate frequency.

Two types of second detector circuits which are commonly employed in radar receiver design are the linear diode rectifier and the plate circuit detector, both similar in form to those employed in prewar television practice. Several factors must be considered in the choice of the second detector operating characteristic. Linear detection of the signal is desirable from the standpoint of realizing the greatest possible visibility of weak radar signals in the presence of noise of comparable amplitude. In the case of lobing radar signals where bearing determinations are made by comparison of the return signal amplitude for two bearing conditions of the antenna radiation, the characteristic of the second detector enters to affect the sensitivity of azimuth response in two ways. The lobing sensitivity is increased by the use of a square law detector, however, the presence of a "jamming" signal of the CW type which may be located off axis can introduce a false bearing indication, which is not present when a purely linear detector is employed. In general, the linear detector has been employed in most radar systems developed during the past war period.

The detailed circuit design of the IF amplifier can be conveniently separated into three quite distinct parts. These include the input circuit design, the interstage arrangement, and the design of the second detector circuit.

2.22 IF *Amplifier Input Circuit Design*

The primary consideration in the IF amplifier input circuit design is the effect of this stage on the over-all receiver noise figure. As indicated previously, this over-all receiver noise figure is dependent on the performance characteristics of the first or input IF amplifier stage to a degree dependent on the loss of the converter stage preceding it. In the military radar field, employing microwave frequencies of 3000 mc and greater, the crystal converter is universally employed and the over-all noise performance is determined largely by the IF input stage. The use of high-gain pentodes in the IF amplifier assures that noise contributions from the following stages are negligible.

Figure 26 represents an equivalent input circuit of the IF amplifier convenient for discussion of the noise performance of this circuit. Here the noise contribution, exclusive of the signal source, is observed to be composed of two sources, one due to shot noise of the first IF stage referred to the grid

circuit and a second noise source related to active grid loading effects.[11]

The optimum IF amplifier input circuit design involves primarily the selection of impedance transformation means which results in the maximum over-all signal-to-noise performance for the radar receiver. It can be shown that this optimum value of impedance transformation is not that value which

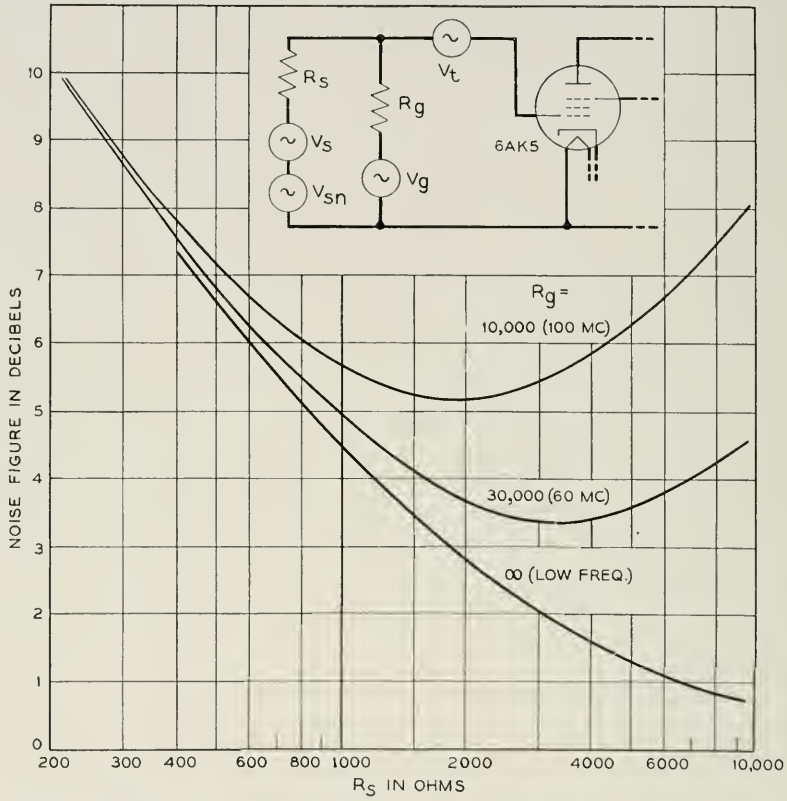


Fig. 26.—Simplified equivalent input circuit of an intermediate frequency amplifier and noise-figure vs signal source impedance.

maximizes the signal but rather is a condition where a definite signal mismatch obtains. This may be understood by an inspection of the characteristics of the two fictitious noise sources illustrated. The shot noise contribution of the vacuum tube employed in the input IF amplifier stage is here represented by the series noise generator V_T. Under conditions of shorted grid this is the only effective noise contribution, and this procedure offers a simple method for determination of the magnitude of this effect.

[11] "Considerations in the Design of a Radar IF Amplifier," Andrew L. Hopper and Stewart E. Miller, *Proc. I. R. E.*, November 1947.

The additional noise contribution under the condition where an impedance is placed in the grid circuit is shown by the noise generator V_G. This noise source is related to the active grid loading. The resistance R_G represents this effective loading which is due to transit time and tube lead inductance effects and, therefore, has a value which is associated with the frequency of operation and the particular design of tube employed. At very low frequencies, $R_G \rightarrow \infty$, the shot noise effects are entirely controlling, and the optimum IF amplifier input grid circuit condition for minimum noise figure is that condition where the impedance of the signal source approaches an infinite impedance. As the frequency of operation is increased, R_G assumes a finite decreasing value and the optimum signal source impedance is given by the relationship illustrated in Fig. 26. It should be noted that the frequency values associated with the above characteristic are indicative of the performance of the 6AK5 pentode, one of the most satisfactory of available tubes for this purpose. It should be observed that the input IF amplifier stage noise figure is independent of the impedance of the signal source providing that a perfect or lossless transformer can be employed to achieve the optinum impedance transformation indicated.

The realization of proper impedance transformation characteristic in the radar IF amplifier input circuit is basically a network problem. The input grid impedance which can be maintained over the desired band of frequencies is limited by the total capacitance present in the grid circuit. For narrow IF band widths, particularly at the lower frequencies, the realizable impedance at the grid may be in excess of the active grid loading value and here the noise performance indicated in Fig. 26 may be realized. However, for wide IF band widths at normal radar IF midband values the maximum grid circuit impedance which can be achieved under the limitation of the grid circuit capacitance will be less than the value associated with the active grid loading and the noise figure obtained will be somewhat higher than the optimum shown. This restrictive design condition is referred to as "band width limited." In modern military radar receivers this condition normally obtains.

To achieve the maximum input coupling network efficiency it is extremely desirable to minimize all parasitic capacitances under control of the designer and to employ the most effective network arrangement available. The double-tuned transformer and autotransformer networks are commonly employed for this purpose. Any loss in the impedance transforming network will result in further degradation of the noise performance, since this loss is effective in reducing the signal energy while having no effect on the tube noise level. The magnitude of the impedance transformation ratio required in a typical radar design case for optimum noise figure is approximately seven times, where a crystal converter of the type shown in Fig. 14

is employed, and the optimum grid impedance at 60 mc as shown in Fig. 26 is approximately 3000 ohms.

The vacuum tube employed in an input stage of an IF amplifier should have the following characteristics to achieve the optimum performance. The tube should be capable of providing sufficient gain to assure that the noise contribution of the following stages is negligible. The input conductance of the selected tube at the intermediate frequency should be as low as possible indicating generally that physical size should be small to assure small transit time effects and low lead inductance values. The noise output of the tube itself should be a minimum, this characteristic being somewhat controllable by proper emission characteristics. The characteristics of the most desirable of the vacuum tubes available to the radar receiver designer during the past war period for IF amplifier purposes is given in Table I.

TABLE I.—*Principal Characteristics of IF Amplifier Tubes*

Type	Heater Power Watts	Plate Current ma	Total Power Consumption Watts	Nominal Transconductance Micromhos	Interelectrode Capacitances Micromicrofarads		Band Merit B_0 mc
					Control Grid to Heater, Cathode, Screen Suppressor Grid	Plate to Heater, Cathode, Screen, Suppressor Grid	
6AC7	2.84	10	4.7	9000	11	5	89
6AG7	4.10	30	9.6	11000	13	7.5	85
6AG5	1.89	7.2	3.0	5100	6.5	1.8	95
717A	1.10	7.5	2.3	4000	4.9	3.8	73
6AK5	1.10	7.5	2.3	5000	3.9	2.8	117

The development of the 6AK5 pentode was the direct result of the necessity for improved radar IF amplifier performance, and details of this development and the performance of this tube have been described elsewhere.[12]

The noise present in a pentode is greater than in a triode, primarily due to the presence of additional grid structures. Because of this fact, a number of attempts have been made to employ triodes in the input stage of the IF amplifier. To prevent oscillation due to large positive feed-back present through the plate-to-grid interelectrode capacitance, neutralization methods have been employed. For moderate IF band widths at 60 mc such experimental designs have shown an improvement in noise figure of slightly more than 1 db over the pentode design; however, the criticalness of the neutralization scheme and the difficulties in extending the performance to wider IF band widths has not allowed this design to be adopted extensively to military radar equipment during the past war period.

[12] "Characteristics of Vacuum Tubes for Radar Intermediate Frequency Amplifiers," G. T. Ford, *Bell System Technical Journal*, Vol. XXV, July 1946.

2.23 Interstage Circuit Design

The design of the IF amplifier interstage circuit is basically concerned with achieving the greatest possible gain per stage while not compromising the frequency characteristic and stability, or complicating the structure to an extent where it will be difficult to realize the designed measure of performance under military operating conditions. The problem can be resolved into the choice of the vacuum tube and the selection of an interstage coupling network.

The gain of a pentode operated into a 2-terminal parallel resonant network which exhibits a maximum impedance R_o at resonance is given by $G_m R_o$ under the restriction $R_o << R_p$. If a band width ΔF is defined as that band of frequencies where the magnitude of the impedance Z of the network is equal to or greater than $\frac{R_o}{\sqrt{2}}$ it can be shown that

$$\Delta F = \frac{1}{2\pi R_0 C}.$$

The gain-band width product of an amplifier stage is a measure of performance of the stage and serves as a criterion for tube performance if a standard load impedance as above is adopted. Then the band merit B_o for this condition will be given by

$$B_0 = \Delta F G_m R_0 = \frac{1}{2\pi R_0 C}(G_m R_0) = \frac{G_m}{2\pi C}.$$

The band merit B_0 of a tube has the dimensions of frequency and may be interpreted as the frequency at which the voltage gain of the vacuum tube is at unity with a plate load impedance restricted solely by the sum of the plate and grid tube capacitances. The ideal IF amplifier vacuum tube will exhibit a high band merit figure, stable operating characteristics over the life of the tube, uniform characteristics during the production period, small size for compact amplifier use and for resulting mechanical rigidity, and finally low-power consumption, desirable from both the power supply standpoint and heat dissipation considerations. The actual types of vacuum tubes generally employed for radar IF amplifiers during the past military program in order of their availability were the 6AC7, 717A, and finally the 6AK5. The improvement in performance achieved through this succession of developments can be observed by reference to Table I and the band merit and power consumption figures for these tube types.

Figure 27 illustrates three types of interstage coupling networks which have been commonly employed in radar IF amplifiers for military purposes. The synchronous single-tuned network design has an advantage of simplicity of construction and permits relatively simple realignment procedures under

the limited resources of military field conditions. This type of interstage has been employed quite widely in radar equipments designed during World War II. To achieve the maximum gain per stage it is necessary to restrict the total shunt capacitance of the interstage circuit to the unavoidable elements due to tube and circuit arrangement. Additional capacitance contributions are avoided by the use of a variable inductance element adjustable through the use of movable magnetic cores to resonate the network to the desired midband IF value. The shunt resistance element is chosen to achieve the desired band width.

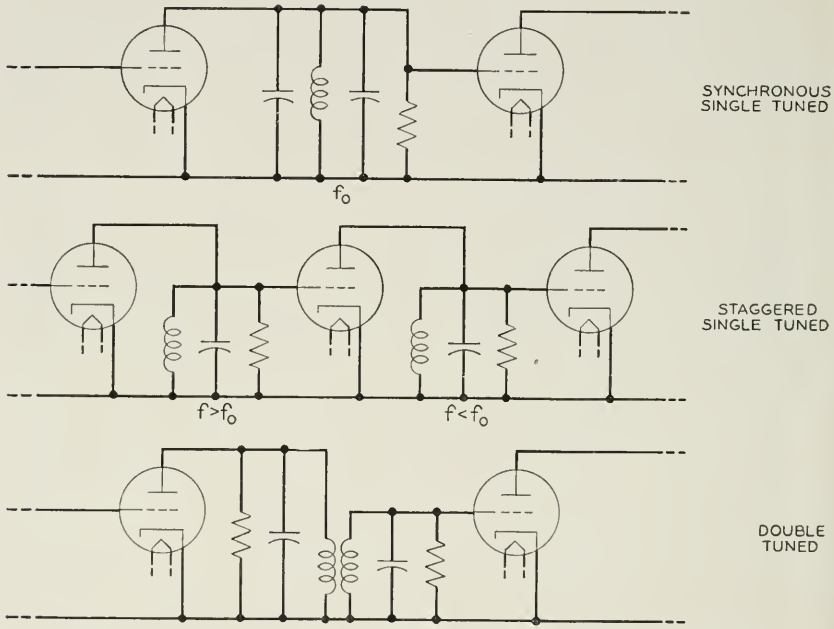


Fig. 27.—Typical IF Amplifier Interstage Circuits. Simplified schematics.

The band width required of each individual interstage circuit of a multistage amplifier of this type to meet an over-all band width requirement of B cycles is given by

$$B = \Delta F \sqrt{2^{1/N} - 1}$$

where ΔF represents the band width of the individual interstage network, defined as the band over which the response is within 3 db of the midband IF value, and N is the number of interstage circuits employed. As the individual interstage band width is increased to achieve the desired over-all value, the gain per stage is reduced and a greater number of stages is required

to meet the over-all gain requirement. The over-all gain of a multistage amplifier employing synchronous single-tuned interstage networks is given by

$$G_0 = \left(\frac{G_m \sqrt{2^{1/N} - 1}}{2\pi C_T B} \right)^N$$

where C_T represents the total interstage shunt capacitance and B is the over-all band width requirement. Table II presents the individual interstage band widths and the maximum over-all gain obtainable for multistage IF amplifiers having a 5 mc over-all band width requirement. Here the use of the 6AK5 pentode is assumed and the total interstage shunt capacitance is assumed to be 12 micromicrofarads. It should be observed that unavoidable misalignment of circuits, aging of tubes, and other such effects all tend to reduce the idealized computed performance under the practical military radar conditions and must be considered thoroughly in the design.

TABLE II.—*Interstage Band Width and Over-all Gain of Multistage IF Amplifiers*

No. of Amplifier Stages	Synchronous Single-Tuned Interstage		Double-Tuned Interstage	
	Interstage Band Width mc	Over-all Gain db	Interstage Band Width mc	Over-all Gain db
1	5	24	5.0	27
2	7.8	37	6.2	47
4	11.5	61	7.6	87
6	14.3	80	8.6	125
8	16.6	96	9.3	162
10	18.7	110	9.8	198

The double-tuned interstage network configuration shown in Fig. 27 is a somewhat more efficient circuit form than the single-tuned variety just discussed, and because of this improved performance has been employed to about the same extent as the synchronous single-tuned type during the past war. Its performance advantage lies in the basic fact that the transmission response curve for this structure has a flat-top characteristic resulting in a slower rate of over-all band width reduction as these circuits are cascaded. The ability to separate the output plate and the input grid circuit capacitances and the elimination of the plate-to-grid coupling capacitor with its additional parasitic capacitance to ground results in a greater gain-per-stage performance. In this structure the resonant frequency of both primary and secondary circuits corresponds to the midband IF and the conditions of equal Q of primary and secondary circuits and critical coupling are assumed. These conditions result in a smooth flat-topped response characteristic having optimum gain performance. The relationship between the individual inter-

stage band width and the over-all band width of a multistage amplifier employing double-tuned interstage circuits is given by

$$B = \Delta F \sqrt[4]{4(2^{1/N} - 1)}$$

which is illustrated in Table II together with the corresponding over-all gain of multistage IF amplifiers of this design.

The third type of interstage network arrangement illustrated in Fig. 27 represents a method employed to realize improved performance of the single-tuned interstage network type by resonating alternate interstages at frequencies above and below the desired midband IF value. This stagger-tuned interstage design permits greater gain per stage together with an increased over-all band width for each pair of amplifier stages over that obtained with the synchronous single-tuned design. In the case of IF amplifiers having six or more stages a variation of this stagger-tuned method can be employed where three successive interstages are considered as a design unit and the individual interstage resonances are adjusted below, above, and centered at the midband IF respectively. To afford a measure of the improved performance of a stagger-tuned IF amplifier we may consider the relative performance of a 6-stage IF amplifier of 5 mc over-all band width employing the 6AK5 vacuum tube. An individual interstage band width of 7.2 mc and an over-all gain of 116 db will result from the use of stagger-tuned interstage circuits while reference to Table II indicates the synchronous single-tuned design would have an over-all gain of only 80 db while the double-tuned design would result in an over-all gain of 125 db. The use of a triple stagger-tuned design would produce a 6-stage amplifier having approximately the same gain performance as the double-tuned example above.

The choice of the interstage network configuration to be employed in a radar IF amplifier must be made considering the circuit efficiency, the gain-frequency stability behavior, and with due regard for the ever-present problem of maintenance of performance under the field conditions of modern warfare. From a standpoint of circuit efficiency alone, it has been shown that the synchronous single-tuned interstage network is decidedly inferior to the more complex forms, but the obvious simplicity of construction of this type and the possibility of adjustment and realignment with simple methods available in the field is a strong recommendation for its adoption in military radar IF amplifiers. The double-tuned circuit has a considerable advantage in circuit performance over the case above, but some portion of this increased efficiency must be sacrificed since it is impractical to construct this network with adjustable elements. Here the normal variations in interstage capacitance with replacement of vacuum tubes or aging effects must not be allowed to reduce the over-all amplifier performance below the design limit. The solution to this problem must be achieved by design of each interstage circuit

to obtain a somewhat wider band under average tube conditions so that under subnormal but acceptable tube conditions the over-all performance of the amplifier is still within requirements. The use of stagger-tuned interstage designs will also result in increased performance over the basic synchronous single-tuned variety, but the maintenance of the over-all performance of a radar IF amplifier of this type involves relatively complex measurements not always possible under military conditions.

2.24 *Second Detector Design*

The final conversion of the IF signal to the video form, as required by the following radar display device, is accomplished by simple detection or rectification. For this purpose either a diode rectifier or a triode operating as plate circuit detector is usually employed. The second detector design follows the practices generally developed for television receivers prior to the war. The diode second detector method has the advantage of simplicity with no plate supply voltage being required, but the performance of such a detector is somewhat limited for the frequencies employed in radar systems. The linearity of rectification of a diode depends on maintaining a high load impedance relative to the internal impedance of the tube. The external load impedance is limited, by the presence of tube and parasitic circuit capacitance and the video band width required, to somewhat less than 1000 ohms for the typical radar case. The internal impedance of the usual available diode is of the order of several hundred ohms so that the linearity of detection suffers. The low value of the diode load resistance is also reflected in the termination of the last IF amplifier stage and affects the gain of this stage.

The plate circuit detector often employed consists of a triode operated near plate current cutoff. Here the detector load impedance is effectively isolated from the plate circuit of the last IF amplifier stage. The linearity that can be obtained from this type of second detector is essentially the same as with the usual diode detector.

The polarity of the detected video output signal may be chosen of either sign by proper circuit arrangement for convenience in the video amplifying and limiting circuits which follow. It is desirable to reduce the amplitude of the IF signal which appears at the output of the second detector to prevent overload and interference in the video amplifier following. This is accomplished commonly by the inclusion of a low-pass filter of simple form in the output circuit of the second detector.

2.25 *Typical Component Designs*

In the military radar system design it has been observed that a maximum video output signal of the order of one volt of noise is desirable as a design

objective resulting in an over-all gain requirement of the order of 110 db. If the radar system employs RF amplification, the entire IF amplification may be provided in one unit. However, in radar systems operating above 1000 mc it has proved advantageous to provide the total IF gain required in two separate amplifier sections. The IF preamplifier assembly is commonly designed to be mounted adjacent to the crystal converter located in the transmitter portion of the radar system and usually consists of two stages of IF amplification. The main IF amplifier is usually located at some distance from the preamplifier, commonly associated with the indicator components of the radar receiver. The main IF amplifier assembly includes the second detector circuit and occasionally one stage of video amplification is included.

The IF preamplifier location as described above is quire desirable, eliminating the need for a long transmission line connecting the IF output circuit of the crystal converter to the input stage of the IF amplifier. As has been discussed previously, the impedance transformation employed in the IF input stage is chosen to realize optimum signal-to-noise performance. The output impedance of the crystal converter is normally of the order of 400 ohms. To assure negligible impedance reflection losses in this circuit, any connecting cable employed would have to be designed to present a characteristic impedance of this order of magnitude which is inconvenient. The practical solution as employed in past military radar systems is obtained by locating the IF preamplifier in close proximity to the converter. The absence of long leads at this IF input stage is also advantageous in reducing the interference pickup into this low signal level point. After moderate amplification the output of the IF preamplifier is usually fed over a 75-ohm coaxial transmission line to the main IF amplifier.

Figure 28 illustrates the converter and IF preamplifier assembly as employed on the AN/APQ-13 and AN/APQ-7 airborne radar bombing equipments operating at 10,000 mc. The local oscillator and silicon crystal converter are arranged in a manner similar to a basic type previously described in this paper. The IF output of the crystal converter is introduced directly into the preamplifier assembly without exposure. This preamplifier is arranged to offer two stages of amplification employing the 717A pentode and using a double-tuned input, interstage, and output network. Figure 29 indicates the circuit arrangement. The gain of this IF preamplifier is 30 db and an IF band width of 6 mc is provided. The output transformer network is arranged to operate into a 75-ohm coaxial transmission line. It should be observed that provision is here included to disable the preamplifier by application of a positive pulse to the cathode circuit of the second amplifier tube. This feature reduces the gain of the IF preamplifier during the short interval coincident with the outgoing radar pulse, which assures that the TR tube

"spike" which precedes conduction will be attenuated and, therefore, less interfering with AFC operation. Further details of this effect will be discussed in a later section of this paper.

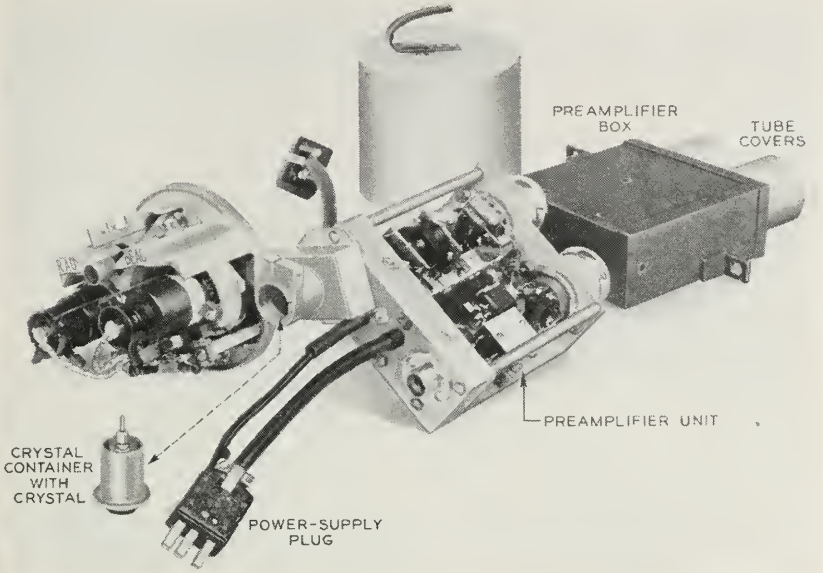


Fig. 28.—Converter and IF preamplifier assembly for AN/APQ-13 and AN/APQ-7 airborne radar bombing equipments.

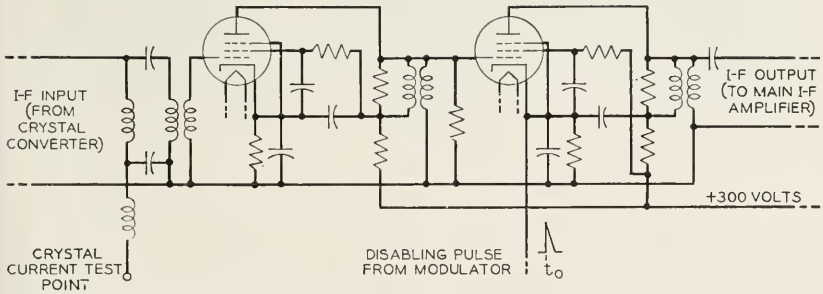


Fig. 29.—Simplified schematic diagram of IF preamplifier component of Figure 28.

Another example of equipment design of an airborne radar converter and preamplifier assembly has been illustrated previously in Fig. 23. In this design the 6AK5 tube is employed and single-tuned interstages and autotransformer input and output networks are employed; however, the general arrangement is quite similar to the design previously described.

The remainder of the IF amplifier gain required is of the order of 80 to

100 db which is usually provided in a main IF amplifier that can be located conveniently within the receiver indicator portion of the radar system. The main IF amplifier is commonly designed as a complete shielded unit, required by the high-gain concentration and desirable from the standpoint of allowing the same unit to be used in several radar systems. Three IF amplifiers are shown in Fig. 30 which well illustrates the technological development in this field during World War II. The first amplifier employing 6AC7 tubes was developed at the beginning of the war, has an over-all gain of 95 db with an approximate band width of 2 mc, and employs synchronous single-tuned

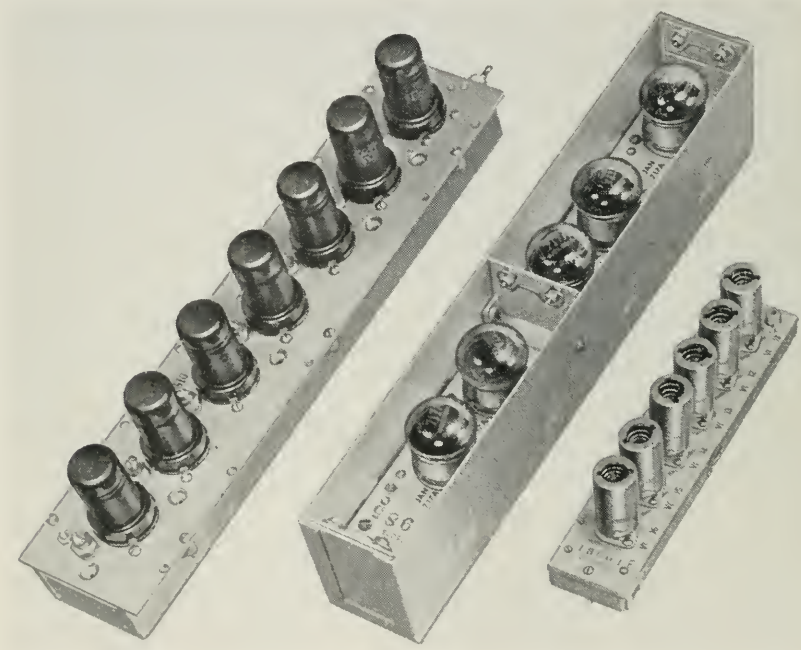


Fig. 30.—Typical IF amplifier equipment designs for military radar applications.

interstage networks. This design was employed extensively in early military radar equipments for land, sea, and air use. It has a total power consumption of 31 watts and weighs 2 pounds 4 ounces. The second amplifier illustrated was developed early in the war primarily for airborne search and interception radar systems and employs 717A pentodes with a double-tuned and single-tuned interstage combination of networks to produce a gain of 85 db with an over-all band width of 4 mc. The total power consumption is 11 watts and the weight here has been reduced to 1 pound 14 ounces. This design of IF amplifier with minor modifications was employed on the major-

ity of airborne bombing radar equipments produced by the Western Electric Company during World War II. The third IF amplifier design illustrated was developed somewhat later in the radar program for specific application to the AN/APS-4 light-weight airborne search and interception radar equipment. This amplifier employs 6AK5 tubes with synchronous single-tuned interstage coupling networks realizing an over-all gain of 100 db for a band width of 2 mc. The power consumption here is 14.5 watts and the weight has been reduced to 9 ounces. This amplifier design has been employed in a number of airborne radar equipments during the later period of the past war.

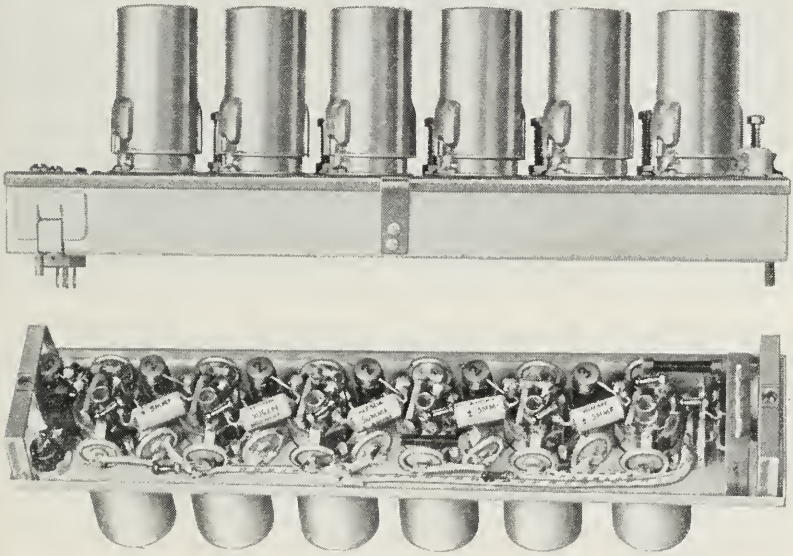


Fig. 31.—IF amplifier design as employed in AN/APS-4 airborne radar equipment.

Figure 31 illustrates further mechanical constructional details of the 6AK5 amplifier described above, and the schematic circuit arrangement is given in Fig. 32. Five stages of amplification and a second detector of a modified plate circuit type is included. A positive polarity video output is obtained from the cathode of the second detector and a single-section video low-pass filter attenuates the IF signal which appears at the detector output. A variable gain control voltage is applied to the plate circuits of the first three stages of the amplifier.

The mechanical arrangement of the components of this amplifier has been devised with a view to achieving optimum frequency and gain stability.

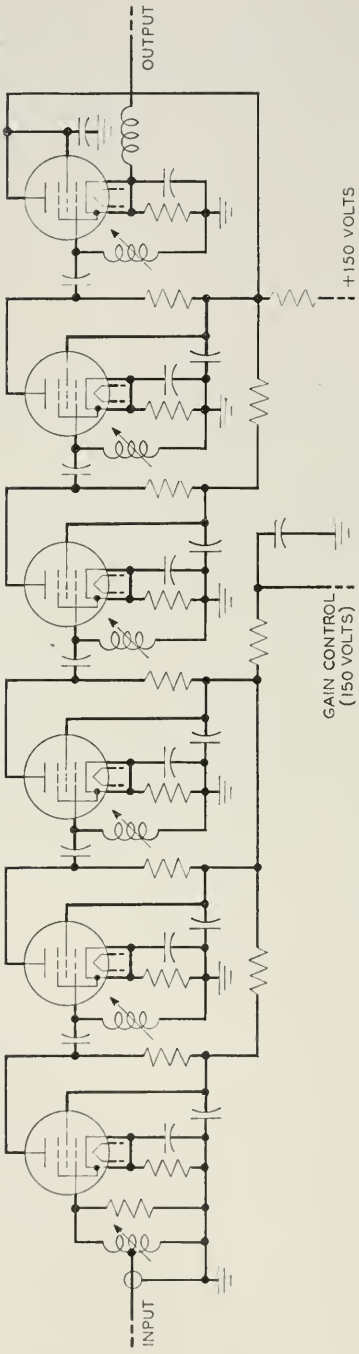


Fig. 32.—Simplified schematic diagram of 60-mc IF amplifier for AN/APS-4 radar.

This equipment design features short and rigid connections and the use of silvered-mica button-type by-pass elements which are mechanically anchored in slots cut into the chassis and soldered in place. The entire unit is arranged to plug into a multipin socket which supplies all power and receives the video signal output. The IF signal input is arranged for plug-in connection at the opposite end of the chassis.

The adjustable inductance elements shown are wound on forms having an approximate diameter of $\frac{1}{4}''$ and the small variation in inductance required to compensate for circuit variations is achieved by the use of tuning screws as illustrated. These coils are adjusted in manufacture by a comparison technique employing factory standards of the same form. The completed amplifier is aligned with mean capacitance tubes and all tuning screws locked and sealed. Sufficient design margin of gain has been included in this design to enable meeting the radar system gain requirements with a complete set of "low-limit" tubes.

2.3 *The Radar Video Amplifier*

The video amplifier of the radar receiver, which follows the IF amplifier and second detector, has as its function the final preparation of the received and detected signal for display. This process involves amplification of the signal in its now video form, introduction of additional coordinate signals and wave forms required for proper display, and often includes modification of the original amplitude characteristics of the signal itself to enhance the presentation. The radar video signal is quite similar in many respects to the television video signal and the circuit technology, therefore, parallels the television art in many respects. Two characteristics of the radar video signal result, however, in somewhat less stringent demands on the radar video amplifier design. The lowest frequency of concern in radar video practice is related to the repetition rate which rarely is found to be less than 250 pps., while it is customary to design television systems to adequately transmit signals of the order of 1 cps. The requirement of faithful reproduction of the radar pulse shape is usually of secondary importance: the quality of presence alone usually sufficing to meet the radar system design objective. In certain fire-control radar systems, however, the radar system band width must be adequate to reproduce the received pulse to an exactness of the order maintained in standard television practice. In general these somewhat reduced transmission requirements for radar purposes result in a desirable economy of circuit elements and power consumption.

2.31 *Gain-Frequency Considerations*

The limiting performance of a video amplifier can conveniently be evaluated by a consideration of the transmission problem at the extremities of the

video band of frequencies. At high frequencies the gain-frequency characteristic of a video amplifier stage employing pentodes as illustrated in Fig. 33 is given by

$$\frac{g_m R_L}{\sqrt{1 + f^2/f_H^2}}$$

where f_H represents the frequency at which the relative gain has been reduced 3 db over the value achieved at the video midband region. This cut-off frequency relationship is similar in form to that encountered in the radar IF

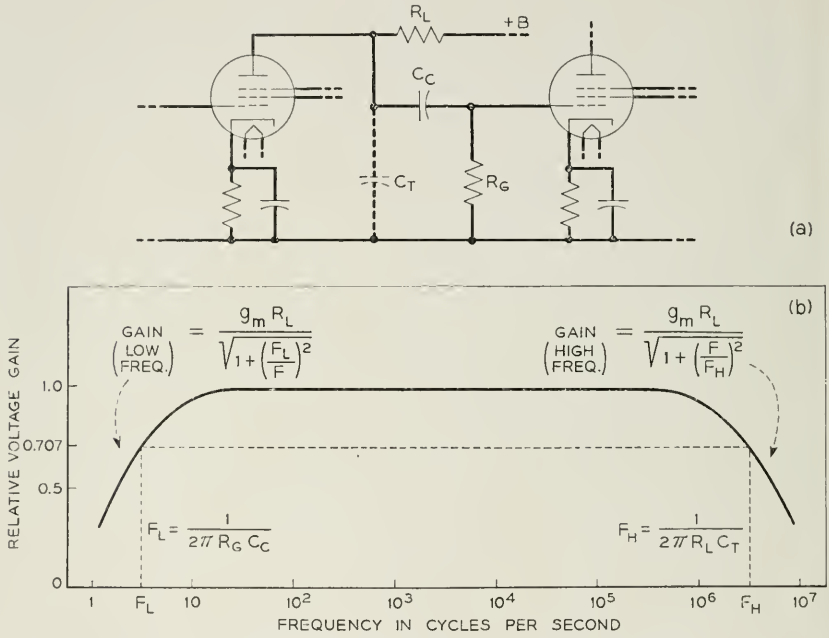


Fig. 33.—Radar video amplifier gain vs. frequency relationships.

amplifier design previously reviewed. In the video amplifier design the vacuum tube band merit B_o again determines the limiting performance of the amplifier, but since the associated video interstage circuit elements contribute considerably to the total circuit parasitic capacitance by reason of their large physical size, the effective band merit of a vacuum tube for video purposes must be considered in terms of the total tube and circuit capacitances. The additional consideration in vacuum tube choice for radar video amplifiers is one of load capacity, since the output signal voltage required for indicator use may range upward to several hundred volts.

In a somewhat analogous manner to the relationships discussed in the

design of IF amplifier interstage networks, the video amplifier performance at high frequencies can be improved by the use of more complex 2- and 4-terminal interstage networks. In military radar systems, however, the added performance realized is more than offset by the undesirability of the additional circuit elements required and the more complex maintenance problems that arise, so that usually only the simple resistance-coupled interstage design is encountered in radar systems.

At the low-frequency extreme of the video band the gain performance of the simple video amplifier is related to the product of the series interstage plate-grid coupling capacitance and the input resistance of the following grid circuit. The low-frequency cut-off of a video amplifier is again defined as the frequency where the gain has fallen 3 db over the value at midband frequencies and is given by $f_L = \frac{1}{2\pi R_G C_c}$. The highest value of R_G that can be employed is related to the grid current characteristics of the vacuum tube chosen. The use of a large value of C_c is undesirable for two reasons. First, the interstage shunt parasitic capacitance increases as a physically larger condenser is employed, which results in poorer high video frequency performance. Second, the use of large coupling capacitances is undesirable from the standpoint of increase in susceptibility to blocking or paralysis in the presence of large signals or enemy jamming. The low-frequency gain response can be improved by certain proportioning of the plate, screen, and cathode by-pass elements also resulting in somewhat less possibility of undesirable feedback through the common power supply impedance.

In certain military radar system designs, multistage negative feedback video amplifiers have been employed. Here considerably greater transmission band width may be realized with the simple interstage network design and an order of improvement in stability results. The feedback amplifier design in these cases usually involves common cathode feedback impedance between the first and third stages.

2.32 *Gain-Amplitude Considerations*

The use of nonlinear gain versus amplitude characteristics in a video amplifier is a condition peculiar to the radar system and represents a considerable departure from established television practice. The factors that indicate the desirability of this treatment of the signal involve the behavior of the amplifier under the extreme range of received radar and jamming signals encountered and the electro-optical characteristics of certain radar indicator cathode-ray tubes.

Definite amplitude limiting of the video radar signal is commonly included in military radar systems. By introducing amplitude limiting at an early part of the video amplifier, complete amplifier paralysis is avoided when

extremely large overloading signals are encountered. These signals may represent strong radar return echoes from objects in the vicinity of the radar antenna or may be due to enemy jamming signals. The ability of the radar receiver to recover in a short time following such serious overloading is an extremely important design consideration. In the case of jamming signals of a continuous-wave or long-pulse form, their effectiveness can be minimized by the inclusion of a high-pass network at the input of the video amplifier.

In the case of radar systems which employ intensity modulated displays of the B, C, and PPI forms the maximum useful brightness that can be attained is limited by "blooming," a phenomenon in which the cathode-ray tube spot on the fluorescent screen undergoes a sudden defocussing when the brightness is increased beyond a critical value. In addition, halation effects are quite pronounced in these long-persistence cascade layer screens and contribute to an undesired masking effect when large areas of extreme brightness are encountered. In these cases it is extremely important to limit the maximum amplitude of the signal that can be impressed upon the indicator. The usual radar video amplifier includes an amplitude limiting stage located early in the amplifier chain whose operating conditions are such as to be driven to plate current cut-off by signals which exceed a preselected amplitude.

The range of useful brightness of the cascade screen radar indicator is severely limited in comparison with the extreme amplitude range of the received radar signals. As has been discussed, the maximum useful brightness has been seen to be limited by halation and defocussing effects while the minimum brightness threshold is controlled by halation and ambient viewing conditions. These limitations of the viewing tube result in a criticalness of adjustment required of the radar operator to achieve the optimum performance of the radar system. In an effort to improve the general reproduction efficiency of the military radar system, several circuit forms have been devised whereby the amplitude of the indicator signal is related to the received radar signal in a nonlinear fashion. In certain instances two parallel amplifier paths have been provided where one path operates in a normal fashion until overload is reached when the second transmission path, designed to properly transmit the higher amplitude signals, becomes effective. In this manner two relatively linear amplification regions are provided with a step or amplitude limiting region interposed. Such a nonlinear circuit arrangement has been referred to as "duo-tone", indicative of the two reproduction regions employed. Another video nonlinear characteristic which was employed in a certain airborne radar bombing equipment developed toward the end of the war was of a logarithmic form realized by a two-path amplifier design. In general this nonlinear treatment of the radar signal amplitude has proven capable of reducing the critical adjust-

ment which has in the past been required of the radar indicator and in this respect has contributed some measure of improved performance under military operating conditions.

2.33 *D-c Restoration Methods*

It is pertinent to examine the exact form of the video signal encountered at the output of the video amplifier as it exists available for indicator use. The presence of series coupling condensers in the video amplifier has removed the d-c component from the signal as detected and, therefore, the average value of the signal is zero. In this form the amplitude of the positive and negative signal excursions are dependent on the form of the signal itself. If such a signal is impressed upon an indicator of the intensity modulated type the average brightness of the scene will remain constant, and the presence of several large amplitude signals will tend to drive any accompanying weak signals below the useful reproduction threshold and effectively fail to reproduce them. In the case of an A-type display where the video signal deflection modulates the beam, the no-signal base line will assume a position on the screen dependent on the video signal form. For these applications it is required that the d-c component be restored to the signal before display. In many other parts of the radar system d-c restoration is required to enable utilizing to the fullest extent the load capabilities of vacuum tubes under the conditions of varying duty cycles of the impressed wave forms. In sweep circuit design a considerable economy of power is achieved by operating the amplifier tubes at or near plate current cut-off for no-signal conditions. Through the medium of d-c restoration, the signal excursions are confined to positive regions only and then regardless of the duty cycle the signal range of amplitude impressed upon the tube is maintained within desired limits.

Figure 34 illustrates three circuit forms which are employed to "reinsert" the d-c component of an a-c video signal. The diode restorer commonly employed in radar systems is shown in Fig. 34a. The impressed input wave, assumed to have an average value of zero as shown, will cause the diode to conduct whenever the signal polarity is negative. During this diode conducting period the condenser C will be charged rapidly, the full effective negative peak signal voltage appearing across its terminals. During the following positive excursion of the signal this voltage difference will be applied effectively in series with the signal. The time constant RC is chosen large with respect to the period of the signal repetition rate and thus maintains this additive bias for the remainder of the signal cycle. Since the effective time constant during the diode conducting period is extremely small valued, limited only by the conductive internal resistance of the diode itself, an extremely small negative excursion time will suffice to

restore the grid circuit to reference zero potential. This particular d-c restorer circuit form is referred to as a positive restorer, indicative of the final polarity of the restored signal. A simple reversal of the diode elements will reverse the polarity of the restored signal.

Figure 34b illustrates the usual radar circuit form of a negative d-c restorer where the diode is eliminated, the normal vacuum tube grid circuit

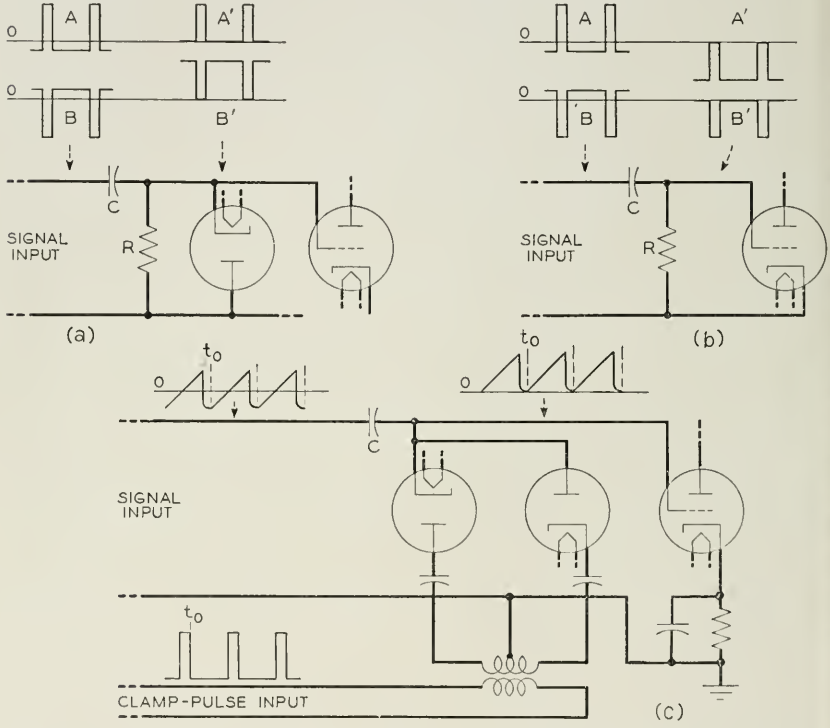


Fig. 34.—Typical D-c Restorer Circuit Forms. Simplified schematic diagrams.

serving here to fulfill this function. Here again an impressed signal form having an average value of zero is assumed, and a negative polarity restored signal is desired at the grid of the amplifier tube. During periods of positive excursions of the input signal the vacuum tube grid will conduct since it is normally operated at zero bias. The series condenser C is accordingly charged relative to the positive signal peak amplitude and this value of potential will be additively combined with the signal during negative excursions in a similar manner to the diode restorer action just described.

A third form of d-c restoration is known as a clamper or synchronized d-c restorer and is illustrated in Fig. 34c. Here a diode bridge circuit

is arranged to be normally nonconductive except during the application of a clamping pulse bias introduced as shown. During this clamping interval, the grid circuit point of the condenser is re-established to reference potential by the low impedance of the conducting diode circuit. At the time of decay of the clamping pulse wave forms the operation of this circuit follows the principles of the d-c restorer types just described. This circuit has been employed less extensively than the preceding simple d-c restorer methods because of the relatively more complex arrangement, but has an advantage in that the impressed signal may be clamped to a convenient reference potential at any particular repetitive point in the cycle.

2.34 *Typical Radar Receiver Video Amplifier Circuits*

The radar receiver video amplifier signal output is required to modulate the indicator by either position or intensity change. In the A type of display the video signal is usually impressed upon a pair of vertical deflection plates of an electrostatic type of cathode-ray tube to present the amplitude characteristics of the signal while the range to the target is displayed as the horizontal coordinate. The maximum video signal amplitude required here to deflect the beam satisfactorily is usually of the order of several hundreds of volts. In the case of B, C, and PPI forms of display the radar video signal is required to intensity modulate the cathode-ray tube. Here a maximum video signal amplitude of 50 volts is commonly required by the radar indicator.

In certain military radar system applications it is desirable to locate the indicator component at some distance from the main radar receiver and video amplifier assemblies. This requirement is commonly encountered in large naval vessel installations where the main radar components may be located below deck and the indicator mounted as a part of the gun pointing mechanism. In such cases video amplifier designs employing video transformer coupling between the output amplifier stage and a coaxial transmission line and between the line and the indicator circuit proper, have proven to be entirely successful.

The development of video pulse transformers for radar purposes represents a considerable advance in the art of communication transformer design. The greatly improved wide frequency band performance of these components is the result of the employment of improved magnetic core materials such as supermalloy having relative permeabilities upward of four times that available in the permalloy materials, improved techniques of coil winding distribution, and the use of additional network elements in the final configuration. Figure 35 illustrates the constructional features of such a video pulse output transformer which has a band width extending

from 100 cycles to 7 megacycles as employed in a naval fire-control radar equipment.

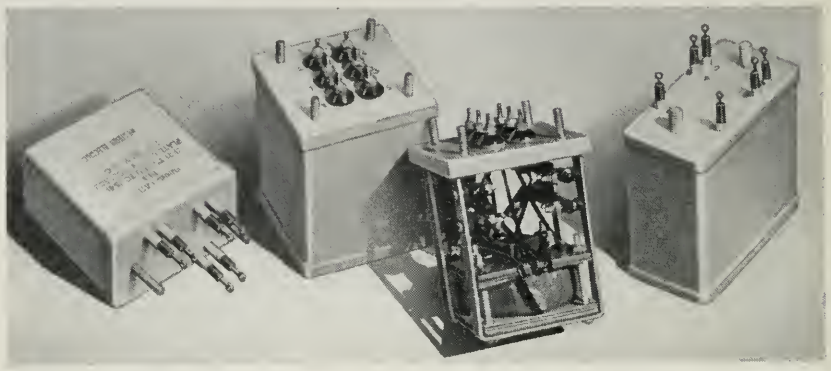


Fig. 35.—Typical designs of radar receiver video frequency transformers.

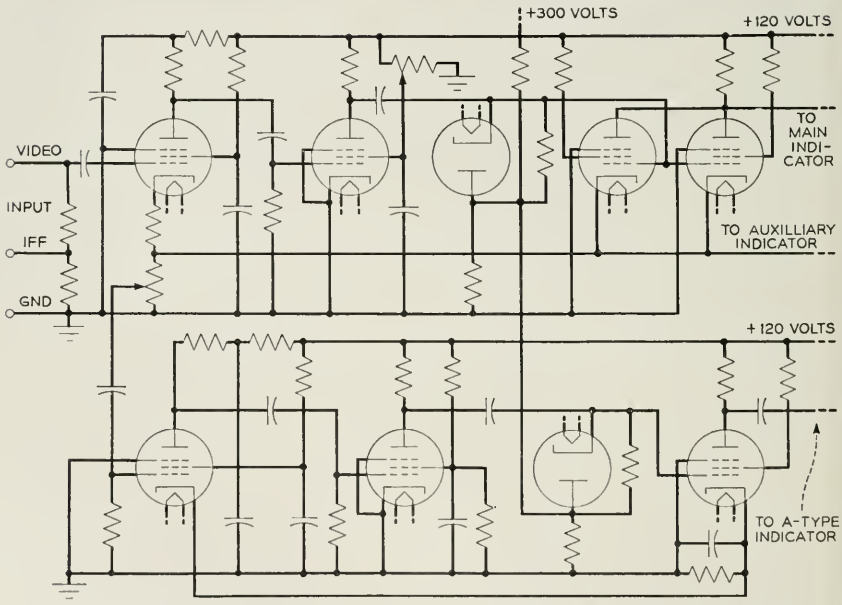


Fig. 36.—Simplified schematic diagram of video amplifiers as employed in AN/APQ-7 airborne radar equipment.

Figure 36 illustrates a video amplifier circuit arrangement as developed for the AN/APQ-7 radar bombing equipment during the latter period of the past war. This system employed two GPI type indicators, one of which was located at a remote station of the aircraft and included also an

A-type indicator which was employed for certain conveniences in operation and maintenance of the equipment. This circuit design includes a main video amplifier for the ground plan indication and a separate amplifier for the A-type display, both of which are of the negative feedback type. The limiting amplifier is included as the second stage with negative d-c restoration included in this grid circuit and diode d-c restoration at the grid of the last stage. To provide sufficient output signal level with the wide video band width required it was necessary to employ two 6AG7 tubes in parallel in the final output stage of the main video amplifier. The local indicator is fed from the plate circuit while the remote navigator's display is fed by means of a low impedance coaxial transmission line. The video gain control is essentially an adjustment of the video amplitude limiting level, the actual signal amplitude being previously adjusted by the IF amplifier gain control. The over-all gain of the main video amplifier is approximately 32 db with a band width of approximately 5 mc. The over-all gain of the A-type display amplifier circuit is approximately 43 db with a useful band width of approximately 6 mc.

2.4 *The Radar Indicator*

The radar indicator assumes a position of extreme importance in the components of the radar receiver. Here with a few specific exceptions all of the electrical information which has been obtained regarding the area under observation is finally correlated and converted into an optical display for use by the radar observer. In the discussion thus far, only the received radar microwave signal properly selected, amplified, and finally converted to the video form has been discussed in detail. The preparation of the additional coordinate and reference data necessary to properly present the complete scene is reviewed in the following sections. In this section the characteristics of the presentation will be reviewed from the standpoint of the requirements imposed by the various radar applications. The electro-optical characteristics of the display device are also discussed.

2.41 *Classification of Radar Display Types*

The number and types of display methods which have been developed for military radar systems during World War II are the result of the varied specific applications to which radar has been subjected. These types of displays are in general related directly to the functional classification of military radar systems previously discussed. It is of interest to consider at this time the various types of indicators which have become common in the radar field. Figure 37 illustrates the basic characteristics of the most important types of radar presentations.

Basically the three coordinates which determine the position of the target

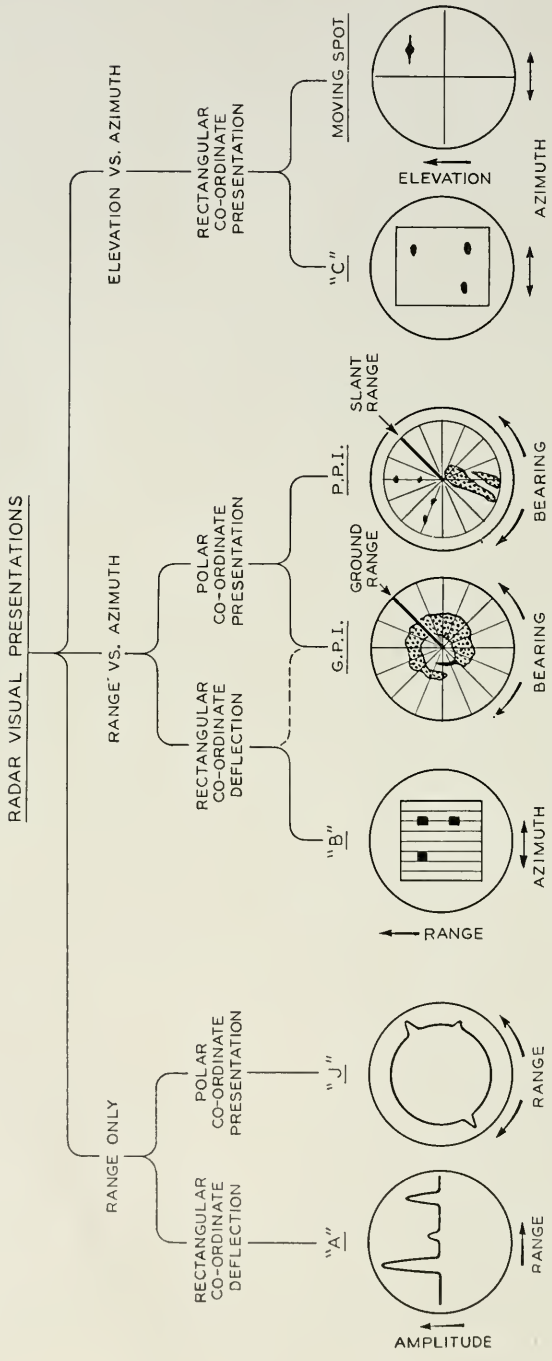


Fig. 37.—Typical military radar display types.

in space and which are determinable from a single radar observation location are the range to the target, azimuth angle with respect to a chosen direction axis, and elevation angle as measured from a convenient reference plane. The classification of radar displays shown in Fig. 37 results from the fact that the only available and convenient display device has the property of resolving only two such coordinates simultaneously. The radar display problem is then one of selecting the most important two coordinates for the specific radar application and choosing the presentation means accordingly. For example, if the radar system under consideration is to be employed on a surface naval vessel against similar naval vessel targets, it follows that elevation angle radar information is redundant and, therefore, type-A or B display patterns are quite satisfactory and are in fact the typical presentations which have been universally employed for surface target fire-control applications. The basic A-type indicator presents range-only data, but for fire-control purposes a modified form is often employed with lobe switching by which accurate training of the radar antenna is possible and bearing information is thus secondarily obtained.

For airborne radar search and bombing applications the presentation is concerned with targets, one coordinate of which is known by other than radar means. Since all targets of interest are in this case located on the ground plane, the relative location of which is determinable by reference to the altimeter and a gyroscopic artificial horizon within the aircraft, it is sufficient here to present all information as a 2-dimensional map. The presence of targets and to some extent their composition is observable as an intensity modulation of the field of view. For this type of application the PPI or its more exact successor the GPI form of presentation is extensively employed.

For military radar applications where fire-control information is desired pertaining to targets which are not confined to a definite plane all three determinable coordinates must be known and, therefore, presented to the radar observer. In certain instances this requirement has been fulfilled by the use of multiple displays each presenting the information regarding one or two coordinates and in cases where gun training is accomplished through separate operators for each coordinate axis, range, bearing, and elevation, this method has proven entirely satisfactory. During World War II the fast moving and highly maneuverable aircraft target has required a more direct and, therefore, faster system of gun pointing. In these cases, the operator has been provided with a display which electronically duplicates a sighting telescope and which merely requires the operator to position the gun (and associated radar antenna) until the target image is centered. To introduce a measure of range to the target the size or form of the target "spot" is often varied in accordance with the range data. For defense against low-level aircraft attacks this admittedly crude range information

has proven quite satisfactory. Similar presentation methods developed for airborne aircraft interception radar equipments have employed, in addition, separate instruments to notify the pilot or gunner at the time when the range to the target was proper for firing of the guns. Another variation in method of obtaining accurate range information simultaneous with the elevation and azimuth data is through the employment of automatic range tracking. In this case after identification and selection of the target has been made and the initial coincidence accomplished, the operator is then free to track in elevation and azimuth with the automatic tracking device continuing to furnish the changing range data to the gun.

Figure 37 indicates the fact that included in these display forms are variations which are a function of the deflection coordinates peculiar to the display device itself. The factors which determine this choice are related to the required form of the presentation from the standpoint of military use, the characteristics of the particular display device available and the mechanical form of the antenna scanning system.

It should be observed that a number of minor variations in the exact presentation is available to the radar system designer within the general classification indicated in Fig. 37. As mentioned previously, the A-type display may be modified to indicate azimuthal pointing errors. In this case, sometimes referred to as a K-type display, the radar system employs an antenna capable of producing two beams of radiation, available one at a time, with azimuthal bearings differing by the order of the beam width. Two signals, each of which is associated with one position of the radiated beam, are displayed in the basic A-type form with one slightly displaced in the range coordinate with respect to the other. By "steering" the antenna until the amplitude response of the desired target appears equal for each image, the target bearing is determined as the direction line bisecting the two antenna lobes.

It is often desirable to limit the display to a small area or to a small selected range interval to enable magnification of this particular portion of the scene. The accurate measurement of range for fire-control purposes can be accomplished on a conveniently small indicator screen by expanding only a selected small range interval of interest. The loss of information at other ranges under these conditions is unimportant. In certain airborne applications it is desirable to present large area information for navigational purposes, but at the time of starting the radar bombing attack the area of interest is limited to a narrow sector extending outward from the plane in the direction of the attack. Here a selected sector may be expanded with a probable increase in accuracy of individual target identification and final bombing accuracy.

In the "range only" classification of Fig. 37 the J-type of display has an

advantage of expansion of the range information by a factor of approximately three times for the same size screen. The A and J types are employed extensively in fire-control radar equipments.

In the range versus azimuth class of radar presentations the B scan historically preceded the other forms shown. Its application was found originally in airborne radar systems for interception purposes. It was commonly employed in conjunction with an auxiliary C-type indicator for target elevation determination. The B type of display suffers from a distortion due to the reproduction of polar coordinate information directly on a rectangular coordinate field. This particular form of distortion is not of major importance where only a few isolated aircraft targets are to be displayed, and in the case of guiding an aircraft to intercept the target the relative expansion of the azimuth scale at short ranges may be a slight advantage. When the B type of presentation is employed for navigation and observation of ground features this inherent field distortion becomes very objectionable when map comparisons of the radar image are required. The B type of display has also been employed extensively on narrow sector rapid scanning naval fire-control radar systems.

The plan position indicator (PPI) type of display was developed to overcome the objectionable distortion of the B-type display and to afford a method of presenting a 360° azimuthal pattern when rotating antenna structures were employed. This form of display essentially replaced the B-type display for aircraft search radar systems and has been since universally employed for ground and naval vessel search systems. Here the linear range trace on the screen is directed outward from the center of the tube, its radial position being synchronized with the instantaneous bearing of the scanning antenna. The map presentation is exact for ground and naval vessel radar locations and for low-flying aircraft radar systems the distortion is negligible, since the slant radar range to the target at low altitudes is essentially comparable to the range as measured on the ground plane. As the altitude of a radar equipped aircraft is increased, the map distortion of the simple form of PPI display also becomes quite objectionable and several modified forms of this display can be employed to improve the presentation. One of these involves delaying the time of start of the linear range sweep by a time interval corresponding to the propagation time of the radar pulse to the ground and return. In this manner a simple but desirable improvement in display is realized. As the military requirements during the later period of the past war became more exacting with the emphasis on high-altitude radar bombing, the remaining distortion of the delayed PPI presentation was found undesirable, and the necessity for accurate map display directly beneath the aircraft resulted in the development of the ground plan indicator (GPI). In this type of display the range trace is deflected as a

nonlinear function of time; its exact time function being dependent on the altitude of the aircraft. The altitude information is obtained from the aircraft altimeter and may be manually or automatically introduced into the radar receiver to produce the proper form of sweep function. These modified forms of PPI presentation were employed extensively in the large bombing through overcast radar program which attained a status of major importance toward the later portion of World War II.

The elevation versus azimuth classification of display forms is essentially restricted to fire-control and aircraft interception radar applications. As previously noted, the C-type display was developed early in the military radar program and has somewhat the same characteristics as the B scan in terms of the distortion which results in the display of polar coordinate data in a rectangular coordinate field without proper mathematical conversion. In the case of aircraft interception radar applications, this type of display is quite satisfactory and has been employed quite extensively for this purpose.

The moving spot (MS) form of radar display is usually associated with a radar system in which conical scanning or lobing is employed. Here the source of radiation of the antenna is arranged and rotated so as to provide a beam whose path describes a cone. If the target is located on the axis of this cone of radiation the signal response will be essentially constant for all instantaneous positions of the beam. If the target is positioned to one side of the cone's axis the received radar signal will be modulated at the frequency of the conical scanning process and the degree of modulation will be related to the angle between the conical axis and the bearing toward the target. This modulation information is utilized within the radar receiver to position an optical image on the face of the indicator screen in accordance with the direction of the target. In radar systems employing this form of indication the observer positions his radar antenna, and accordingly the associated weapon, to center the target image on the indicator. Mechanical or electronic cross hairs are employed as the reference axis. A measure of range to target information is often introduced into this form of display by assigning an arbitrary but distinctive size or shape to the target spot which can be varied in accordance with the range to the target being observed.

2.42 *The Cathode-Ray Tube*

The cathode-ray tube is without serious competition as the ideal radar indicator, primarily because of its unique high-frequency electro-optical response characteristic. Since the radar presentation requirements are not unlike those encountered in television practice, it is natural that the cathode-ray tube development for radar purposes should have progressed along simi-

lar lines originally established by prewar television. Two general types of cathode-ray tubes have been commonly employed in the military radar program, electrostatic and magnetic, these classifications being indicative of the deflection and focussing method employed.

Electrostatic Deflection Type

A typical form of electrostatic-type of cathode-ray tube suitable for radar indicator purposes is shown in Fig. 38. In this tube type the electrons emitted from an indirectly heated cathode surface are initially formed into a beam by passage through an aperture which serves as a beam density or ultimate brightness control element. Following this, the electrons proceed through another aperture which is maintained at a positive potential with respect to the cathode. This first anode together with the following second anode forms an electron lens system which focuses the beam on the fluores-

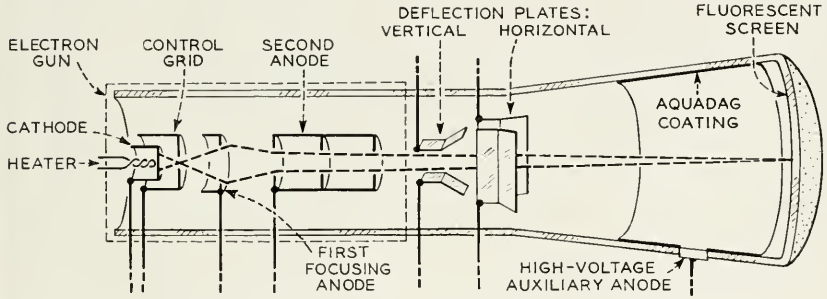


Fig. 38.—Schematic diagram of an electrostatic type cathode-ray tube.

cent screen surface. The relative potentials of these elements serve to enable focussing of the beam by electrical means. The deflection of the beam for scanning purposes is here accomplished by the introduction of an electric field formed by the application of potential across the deflection plates shown. Two pairs of plates enable separate horizontal and vertical deflection to be employed. The plates are formed as shown to enable obtaining the maximum deflection per unit of electrical potential applied without interference with the beam under large deflection conditions. The high-voltage auxiliary anode is provided in certain tube types to further accelerate the beam after deflection without an appreciable reduction of deflection sensitivity and results in an image of increased brightness.

To realize optimum performance of an electrostatic-type cathode-ray tube several precautions must be observed. Serious defocussing of the beam as it is deflected will result if the average potential of the pair of deflection, plates is allowed to vary substantially from the value present at the second anode. To minimize this effect, balanced sweep deflection amplifiers are

commonly employed in radar receivers employing electrostatic deflection, and the second anode is maintained at the average potential of these deflection plates. Astigmatism, the selective focussing in one direction on the screen at the expense of focus in the other, results from the mechanical limitations whereby the electric fields of the two pairs of deflection plates cannot be made effective at the same point within the tube. Some improvement can be realized in this respect by operating the pairs of deflecting plates at slightly different average potentials.

The limitation in deflection response at high frequencies is a function of the total deflection circuit capacitance. To eliminate the blocking condensers with their considerable parasitic capacitance to ground it is common radar indicator practice to operate the deflection plates directly from the

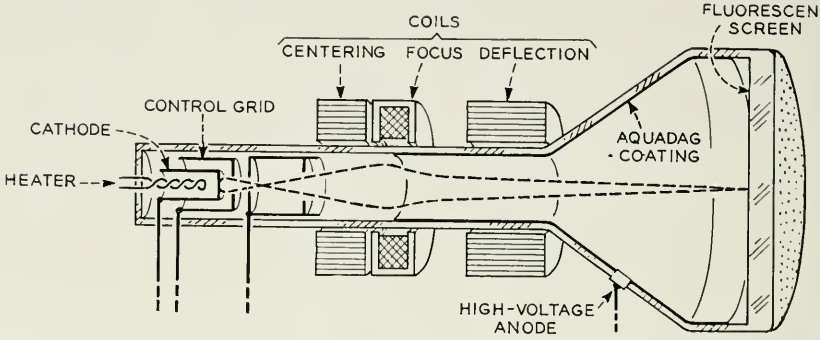


Fig. 39.—Schematic diagram of a magnetic type cathode-ray tube.

plates of the sweep amplifier tubes and then accordingly operate the cathode of the cathode-ray tube at a high negative potential.

Magnetic Deflection Type

A typical form of a magnetic-type cathode-ray tube is illustrated in Fig. 39. In this tube the electron gun structure comprises a heater, cathode, control grid, and second or screen grid which, together with the second anode, usually formed by an aquadag coating within the glass envelope and which is maintained at a high positive potential, roughly delineates the beam. The second grid in this case serves to shield the control grid from the high potential second anode and results in an improved control grid characteristic. The beam of electrons is focussed through the use of a magnetic field located as shown in Fig. 39, and produced by direct current flow through a coil or by a permanent magnet structure. The centering of the beam upon the screen under no deflection conditions is accomplished by the use of a distorting field commonly introduced as a part of the deflection coil and focussing assembly. The deflection of the beam for a rectangular coor-

dinate display system is here accomplished by magnetic deflection fields perpendicularly disposed and produced by a pair of deflection coils located at the junction of the neck and bulb as shown. In the polar form of display (PPI) usually only one deflection coil is employed for the production of the radial sweep, the angular deflection being produced by rotation of this coil about the neck of the cathode-ray tube in synchronism with the rotation of the antenna.

It is of interest to compare the relative characteristics of the electrostatic and the magnetic-type of cathode-ray tubes from the standpoint of their application to radar indicators. The electrostatic-type of cathode-ray tube has a distinct advantage of lighter total weight for the tube itself and the associated deflection circuits required, which in the case of airborne radar equipment design is an important factor in its favor. In general, it is desirable to present a large radar display field. For the larger diameter cathode-ray tubes the magnetic-type tube has an advantage of shorter over-all length which has proven an important equipment design factor for airborne radar equipment where the available operating space is severely limited. The magnetic-type cathode-ray tube requires weighty focussing and deflecting assemblies and large power supplies to furnish the heavy deflection current required, but because of the higher anode voltages which may be employed here, the screen brightness achieved is considerably greater than that available in the usual electrostatic type of cathode-ray tube. If the anode potential of an electrostatic type of cathode-ray tube is increased and hence the screen brightness, the deflection sensitivity is seriously impaired and a difficult deflection amplifier design results. From a deflection point of view, it is possible to achieve somewhat better performance in reproducing extremely high-speed sweeps by employing an electrostatic-type of indicator which has somewhat less serious parasitic elements which act to limit the high-frequency response. The final choice of type of cathode-ray tube for the radar indicator is dependent on the specific detailed considerations of the system in hand. No general and fast rules governing this decision are evident.

Characteristics of the Fluorescent Screen

The fluorescent screen of the cathode-ray tube upon which the final radar information is converted into the desired visual form consists of a deposit of certain materials which exhibits fluorescence when bombarded with a high-velocity electron stream. Phosphorescence, the continued emission of visible light after bombardment has ceased, is also a property of all of these screen materials. These screen materials, referred to as "phosphors", have characteristics dependent on their physical form as well as their chemical composition.

Two general types of phosphors have been commonly employed in military radar systems classified according to their phosphorescence characteristics. The medium persistence class of phosphors exhibit decay times of the order of milliseconds and are composed of a single layer of Willemite (zinc orthosilicate). This type of screen exhibits a green luminous response and is employed extensively in reproducing high-speed wave forms such as encountered in high-speed scanning systems, and for general A-type presentation purposes. The visible light output decays to the order of 1% of its initial value in approximately 50 milliseconds after electron excitation ceases. For photographic purposes other phosphor compositions of the same persistence class whose useful light output has a higher actinic value are commonly employed.

The long-persistence phosphors are composed of two layers of screen material which combination exhibits sustained phosphorescence, the visible light output decaying very slowly after cessation of bombardment. The double layer or cascade screen consists of an innermost layer subject to the direct influence of the electron stream which is composed of a silver activated zinc sulphide and a second layer adjacent to the glass envelope which consists of copper activated zinc cadmium sulphide. The first-named material fluoresces with an extremely brilliant blue light under bombardment and exhibits a rather rapid decay characteristic. The second layer is in turn excited by the blue radiation from the first layer and responds with a yellow visible emission which persists for a matter of several seconds after excitation ceases. The initial blue flash is appreciably absorbed by the second phosphor layer, but usually further optical attenuation is required to prevent eyestrain and degradation of night vision of the observer. This is commonly provided by the use of an amber optical filter placed over the screen face. The long-persistence characteristics available in this type of tube have proven invaluable in military radar systems which feature slow antenna scanning. In many of these systems the time between successive scans of the target may be of the order of a second or more and only through the use of the cascade-type long-persistence tube can the image be retained for this period of nonexcitation. Another property of the long persistence class of cathode-ray tube screens which is of advantage for radar purposes is an accumulative increase in brightness with successive scans of the target. Since the target image is usually repetitive as regards position on the screen, the image brightness will increase with successive scans while because of the random character of noise no such increase in noise image will result, and a small but evident signal-to-noise improvement obtains. The long-persistence type of screen characteristic is employed in the majority of military radar indicators of the B, C, and PPI-types.

Another general characteristic of the cathode-ray tube screen which influ-

ences the over-all system performance is the range of useful brightness available. The extreme variation in the radar response of targets in an area under observation has been discussed previously. The inability of the cathode-ray tube screen to convert this extreme range of electrical signals to a correspondingly large optical brightness range has been a restriction on the performance of military radar systems. Isolated measurements of the useful brightness range available in a cascade screen cathode-ray tube indi-

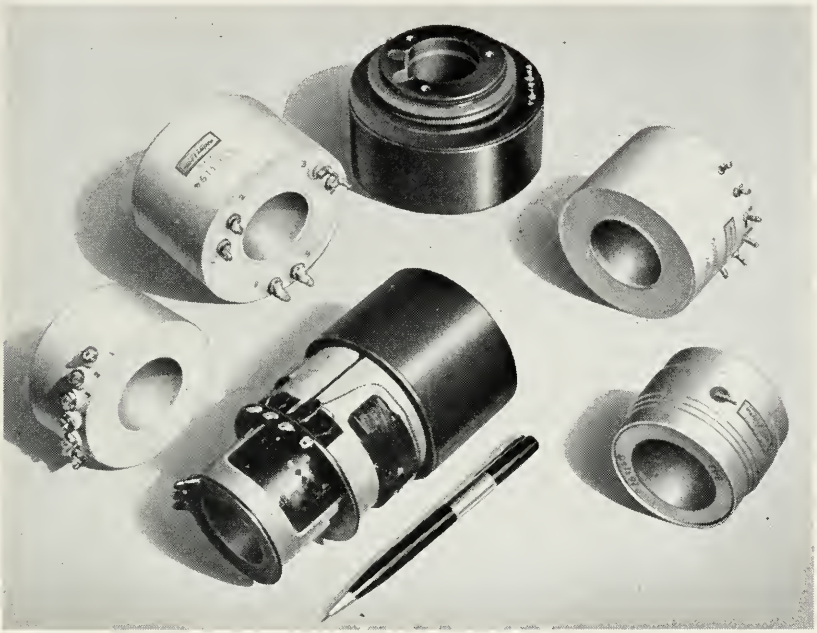


Fig. 40.—Typical magnetic focus and deflection coil designs as developed for military radar purposes.

cates that it is of the order of 10 to 1. The development of the logarithmic video amplifier and "duo-tone" is an attempt to improve this situation. In general this limited useful brightness range of the radar indicator results in a critical adjustment of the operating region of the indicator tube. In a practical military radar system operating under wartime conditions the inclusion of such a critical adjustment always effectively results in a reduction of performance from the optimum achieved in the laboratory.

2.43 *Typical Radar Indicator Component Designs*

Figure 40 illustrates a few typical component designs of magnetic deflection and focus coil structures developed for specific military radar applica-

tions during the past war. The deflection coils illustrated include both air and permalloy core structures as used in rectangular and polar types of displays. Where the radar system involves extremely short time interval sweep wave forms, the maximum inductance which can be employed in the deflection coil is limited by the power supply voltages available, and in these indicator designs the air-core type of deflecting coil is usually employed. Where the sweep wave form is relatively slower the permalloy core types have been extensively employed with an effective improvement in deflection sensitivity. For the PPI form of display a toroidal coil structure has been devised which contains two distributed windings connected in an opposing sense. The internal leakage flux of such a structure is essentially uniform and, therefore, satisfactory for magnetic cathode-ray deflection purposes. The usual PPI type coil structure is arranged to mount within a large ring ball bearing to enable rotation around the neck of the tube with provisions being included on the coil housing for slip rings to afford connection to the deflection coil proper.

With the increased emphasis on extremely accurate radar presentations, which developed during the later war years especially in connection with the radar bombing program, the design and manufacturing tolerances allowable in connection with the large scale production of these magnetic deflection coils were severely reduced. Figure 41 illustrates the constructional details of a deflection coil as employed in the AN/APQ-7 radar bombing equipment. The presentation in this instance is of the GPI type employing rectangular coordinate deflection and extremely fast sweep wave forms. This deflection coil structure employs accurately formed open air-core windings which are initially adjusted and cemented to concentric phenol plastic cylinder forms. This design features a vernier rotation adjustment of the horizontal and vertical pairs of coil assemblies to meet a manufacturing scanning requirement of $90° \pm 0.5°$.

Two examples of focussing and centering structures for magnetic-type cathode-ray tube radar indicators are included in Figure 40. The focus coil consists of a simple winding located axially about the neck of the tube with a shielding magnetic structure containing an annular air-gap which restricts the external field to a region including the cathode-ray tube electron beam. This structure is designed to produce a uniform magnetic field distribution in the complete area of the beam to avoid defocussing effects. In certain early-design airborne radar equipment applications where the equipment was subjected to extreme variations in ambient temperature over short periods of operation some difficulty in maintaining optimum focus was experienced. This defocussing, due to the change in coil resistance with ambient temperature and in part to dissipation in the winding proper, is minimized in the designs shown by the introduction of a varistor element

mounted in close proximity to the winding whose resistance temperature characteristic was chosen to compensate for similar characteristics of the winding proper. The beam centering structure included in the designs shown employs two pairs of coils arranged upon a closed magnetic core structure adjacent to the focus winding. The perpendicularly disposed magnetic fields are produced by direct current in the same fashion as discussed in connection with the deflection coil assembly.

Figure 42 illustrates the operation of a permanent magnet type of focussing and centering structure developed during the war and which was em-

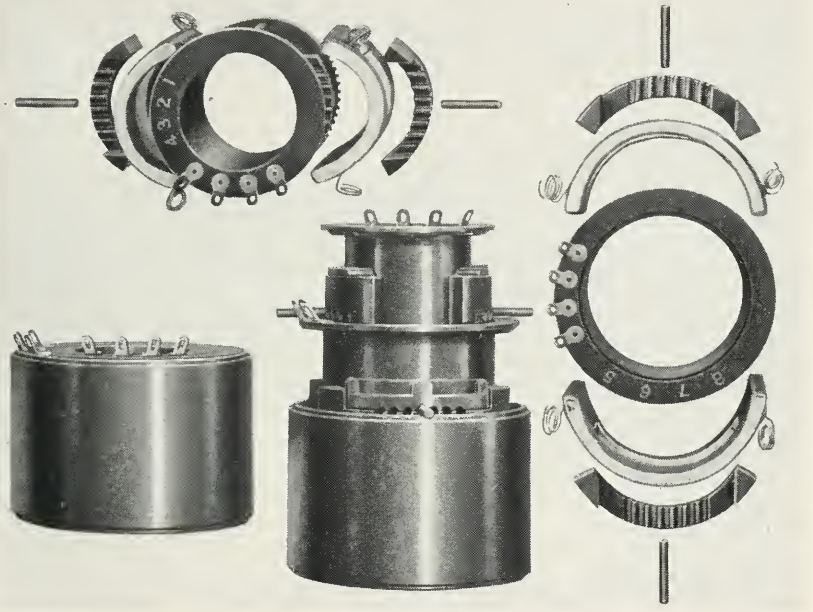


Fig. 41.—Constructional details—deflection coil design for AN/APQ-7 radar bombing equipment.

ployed extensively in airborne radar applications. Here a permanent ring magnet is equipped with a variable internal mechanical magnetic shunt whereby the focussing magnet field can be varied as shown. The centering of the beam is accomplished by means of a centering ring located as shown capable of being mechanically controlled along two perpendicular paths. Mechanical linkage arrangements provide location of the focus and centering adjustment controls on the indicator panel convenient to the operator. The permanent magnet type structure has the advantage of maintenance of proper focus and centering under conditions of extreme variations in ambient temperature. Similar permanent magnetic type structures have been em-

ployed to permanently deflect the beam for off-center displays such as the B-type and certain modified sector scanning PPI forms. Here the ampli-

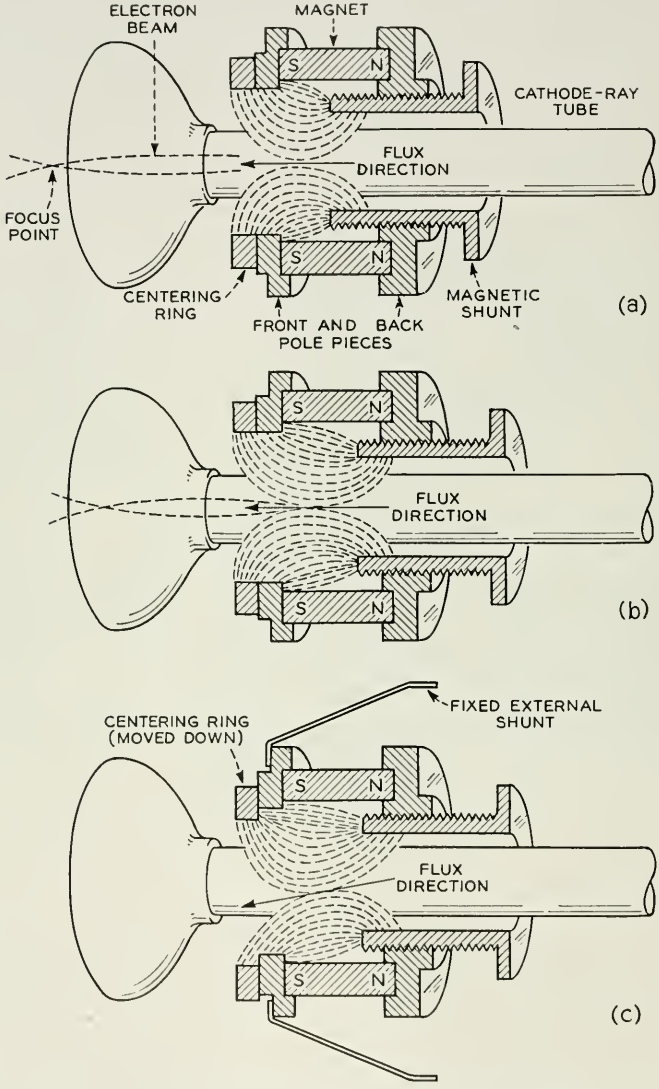


Fig. 42.—Operational diagram of permanent magnet type of focussing and centering structure for radar indicator.

tude of deflection required is approximately equal to the radius of the screen and this precludes the use of the usual beam centering structure located near the gun structure portion of the neck of the tube. To prevent physical

interference by the neck of the tube with the deflected beam such off-center display deflecting structures are mounted close to the junction of the neck and bulb of the cathode-ray tube.

As indicated previously the blue-flash characteristic of the excitation of the first layer of the cascade-type screen is usually reduced in intensity through the use of an optical filter placed between the screen of the cathode-ray tube and the observer. These optical screens are usually constructed of an amber-colored transparent plastic whose particular optical transmission characteristics are chosen in accordance with the particular phosphor and speed of scanning employed. It is common practice to engrave general range and direction reference lines on such screens, and in many cases variable edge lighting of these engraved screens is employed to enhance the display.

In certain applications of radar, namely, airborne reconnaissance and bombing operations, it is desirable to obtain a photographic record of the display to be used for training, briefing, or damage assessment purposes. It is customary in these cases to employ a stop-frame moving picture camera attached to the indicator and exposed periodically as desired. Figure 43 indicates the design of a radar indicator viewing attachment which was employed in connection with an airborne bombing radar equipment toward the end of the past war. In this design a partially silvered mirror is located at a 45° angle with respect to the axis of the cathode-ray tube. An illuminated image of an adjustable course marker located below the mirror is observed superimposed upon the radar presentation with negligible parallax distortion. A portion of the light of the radar image is also reflected from the surface of the partially silvered mirror, and together with the direct image of the course marker is available for photographic recording by the camera mounted as shown. Automatic exposure at any preselected time interval is provided, the exact exposure time being controlled by impulses derived from the azimuthal scanning mechanism of the radar antenna. The photographic recording of radar displays has become a matter of prime importance in the modern military operations and has added another consideration for the military radar indicator designer.

Figure 44 indicates an equipment arrangement of a PPI indicator as employed in the AN/APQ-13 radar bombing system. In this system the indicator is designed for convenient overhead mounting in the restricted radar operating space available in modern bombing aircraft. The deflection coil in this equipment is rotated about the neck of the 5″ diameter long-persistence cathode-ray tube by a geared selsyn motor energized from a similar selsyn unit mechanically linked to the rotating radar antenna. Permanent magnet focussing and centering is employed in this particular indicator. Figure 45 illustrates a somewhat similar packaging arrangement for

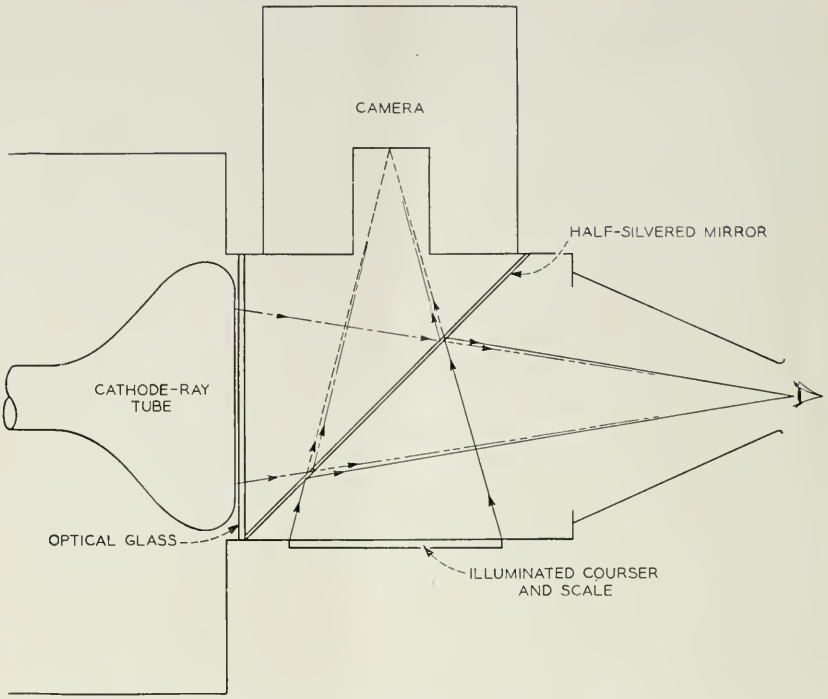


Fig. 43.—Mechanical arrangement of a radar indicator viewing and photographic attachment for AN/APQ-7 radar bombing equipment.

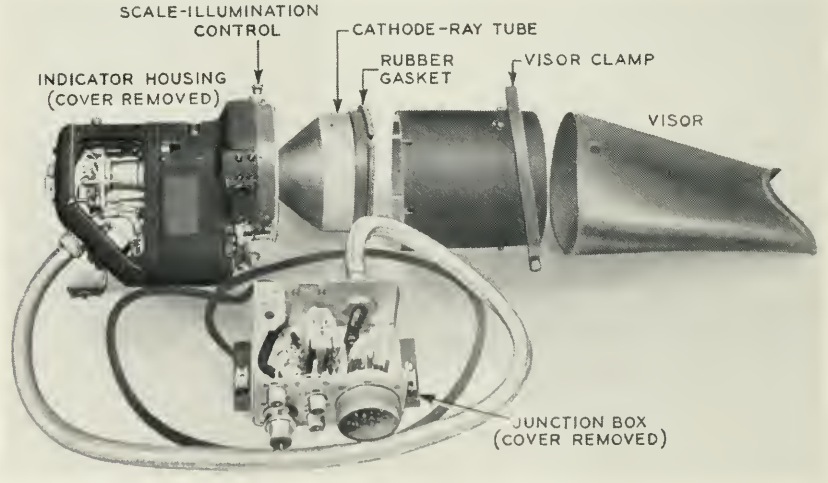


Fig. 44.—Indicator design for AN/APQ-13 radar bombing equipment.

an airborne radar low-altitude bombing equipment as employed extensively during the past war. Here the display employed is of the B variety utilizing a 5″ long-persistence cathode-ray tube. The focussing and centering magnetic fields here are produced by coils located as shown. The azimuth sweep voltage is obtained by means of a potentiometer mechanically geared to the sector scanning antenna. The permanent off axis deflection of the zero range line of the B display is here obtained by the use of a permanent magnet yoke mounted at the junction of the neck and bulb of the cathode-ray tube.

Figure 46 illustrates a typical form of an A-type indicator as developed for the SH Naval and Mark 16 mobile fire-control radar equipments. This unit includes provision for receiver tuning and video limiting adjustment convenient to the observer. The SH and Mark 16 radar systems employ

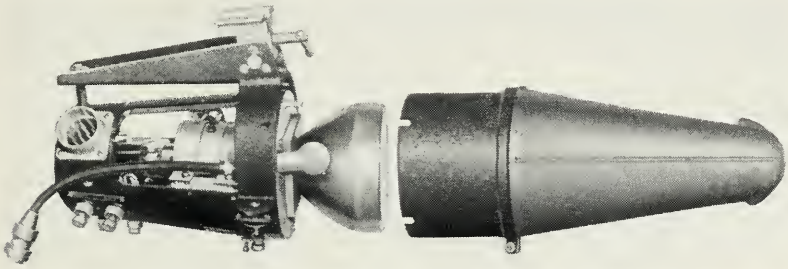


Fig. 45.—Mechanical design of AN/APQ-5 type B radar indicator.

both lobing of the antenna for precise azimuth bearing determination and also continuous rotation of the antenna for general search purposes. The indicator shown in Fig. 46 is employed when the lobing system is in operation and features a precision range sweep as well as the full range display. The equipment design here reflects the severe mechanical requirements which must be satisfied for electronic equipment which is to be employed on naval vessels or for trailer mobile ground service.

2.5 *The Radar Sweep Circuit*

2.51 *Function*

The sweep circuit components of a radar receiver are required to generate the specific voltage or current wave forms necessary to properly display the radar-received information in the desired form. These wave forms must also be actuated by or related to the various coordinate or computed types of data furnished to the radar receiver. The actual detailed configuration of the circuit employed for this purpose is dependent on the form of the input

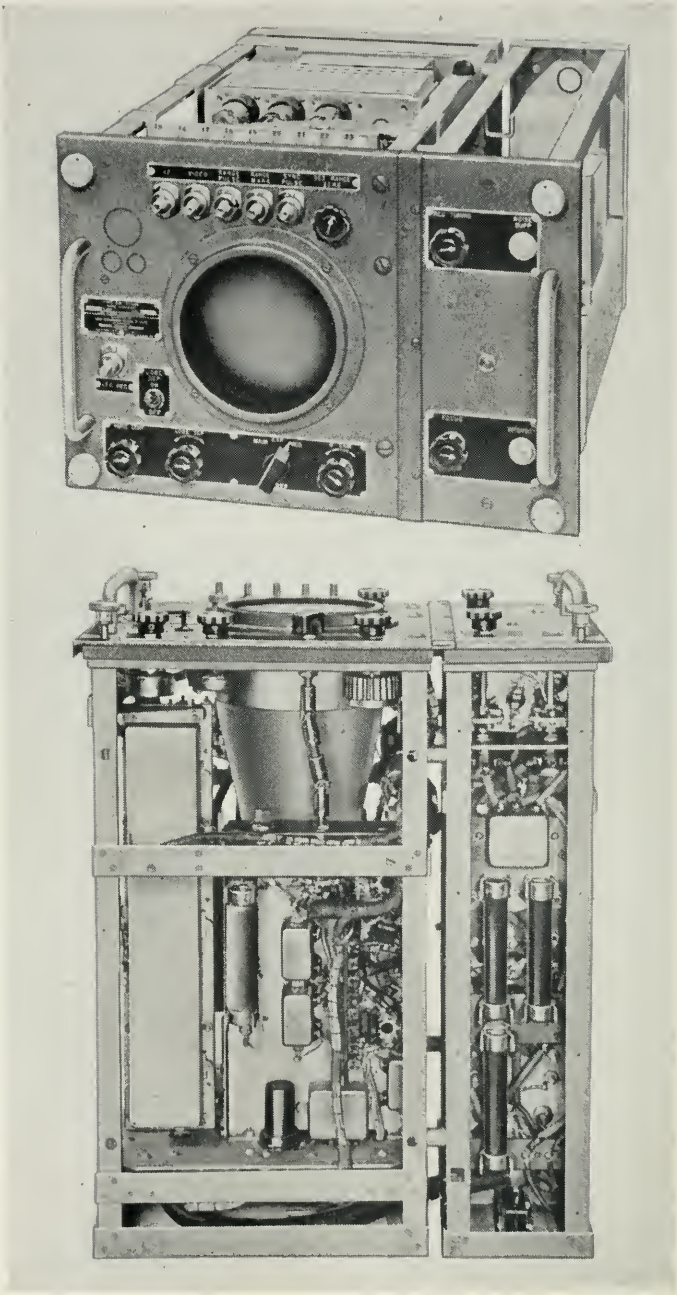


Fig. 46.—A-type indicator design for SH Naval radar equipment.

information available and upon the indicator deflection methods to be employed.

The sweep deflection problems may be separated into two quite distinct categories dependent on the speeds of operation involved. The slow-speed deflection systems originate with the required deflection of the beam of the indicator cathode-ray tube in accordance with the instantaneous position of the axis of propagation of the antenna structure. Since the antenna structures with which we are concerned at radar frequencies are of comparatively large dimensions, their motional velocities are extremely limited and accordingly the electrical information describing these slow mechanical changes contains only low-frequency components. The deflection problem associated with this information, in general, offers little difficulty to the radar receiver designer. Commonly employed methods of slow-speed sweep deflection include the use of potentiometers, selsyn generators or variable capacitors mechanically linked to the deflection axes of the antenna structure and associated circuits of relatively simple form whereby the electrical changes in the characteristics of these devices are more or less directly impressed upon the proper deflection axis of the indicator. In the case of the PPI form of display, it is quite common to synchronize the rotation of the deflection coil about the axis of the cathode-ray tube with the azimuth bearing of the antenna through the use of selsyn motors or servo mechanisms. In general, slow-speed sweep deflection problems are associated with bearing coordinate data only.

The determination of range to the target, on the other hand, requires high-speed scanning whereby the time interval encompassing the time of propagation and return of the radar pulse over the selected range interval must be completely displayed upon the indicator screen. The total time interval available to deflect the beam for range measurement purposes may extend from 2500 microseconds which represents a range measurement of approximately 240 miles down to perhaps 6 microseconds representing an expanded interval of approximately 1000 yards useful in certain fire-control applications. Here, with extremely small times available for deflection purposes, the radar receiver designer is faced with difficult circuit design problems where the usual negligible parasitic circuit elements now severely restrict the circuit performance. In the following discussion, therefore, emphasis will be placed upon the design factors involved in the development of high-speed radar sweep wave forms.

The radar sweep circuit can be considered as providing the following functions:

1. Generation of time wave forms.
2. Generation of display sweep wave forms.
3. Amplification of sweep wave forms.

2.52 *The Timing Wave Form Generator*

The generation of the timing wave forms consists of the preparation of specific voltage wave forms for use in the following sweep generator in accordance with timing information available at the radar receiver input. These timing data may consist of a pulse coincident or related to the outgoing radar microwave pulse serving as a reference for range display or, in the case of direction sweeps, may consist of signals related to the instantaneous position of the antenna.

The basic wave forms employed in this connection consist of rectangular pulses where the duration of the pulse may be controlled to serve as a measure of time, extremely short-duration pulses useful as time markers, and various combinations of these. In general, these wave forms are characterized by their nonsinusoidal form. The generation of these nonsinusoidal wave forms is accomplished by a number of specialized electronic circuits, which though apparently quite complex can be resolved generally into a combination of relatively simple basic circuit forms.

The Multivibrator

The "trigger" or multivibrator circuit was developed nearly thirty years ago and provides the fundamental circuits for the sweep circuit designer. Figure 47a illustrates a simple historical form of a trigger circuit which is of the Eccles-Jordan type. The essential current-voltage relationship which characterizes this circuit and all circuits employed for this purpose is a negative resistance characteristic which exists over a limited portion of the operating range of the device. In the case of the electronic circuit shown in Fig. 47a this negative-resistance characteristic is bounded by two stable limiting conditions. Referring to the trigger circuit of Fig. 47a, the chronological order of operation can be described as follows: Assume V_1 is conducting a somewhat larger current than V_2 so that the potential at the plate of V_1 is lower than the corresponding point at V_2 due to the voltage drop across the plate resistor R_1. This condition further implies that the grid potential of V_2 as determined by the connection from the plate of V_1 through the coupling resistor R_3 is lower than that at the grid of V_1. Similarly the grid potential of V_1 is at a higher positive potential, due to its connection with the plate of V_2. The action is cumulative and results in stablizing the circuit under the condition where the plate current of V_2 is entirely cut off and the voltage drop across V_1 is less than the grid bias voltage E_c.

If now a voltage is impressed across the input terminals of either a positive or negative form, the circuit will be driven away from this stable equilibrium condition as follows. Assume now that a large positive pulse be applied to the circuit shown. The tube V_1 which is operating in a conducting con-

dition will not be affected but the grid of V_2 will be raised in potential by the amplitude of the enabling pulse. The plate current flow in V_2 under this influence will reduce the plate potential of V_2 and accordingly will tend to decrease the positive bias of V_1. The accompanying plate current reduction of V_1 will increase its plate potential and this will result in increasing the grid potential of V_2 through the coupling resistance R_3. Again the cumulative effect will be to abruptly cut off the plate current of V_1 and operate V_2. Thus an abrupt switching of this electronic circuit results when a single enabling pulse is impressed upon it. The wave form across one of the plate

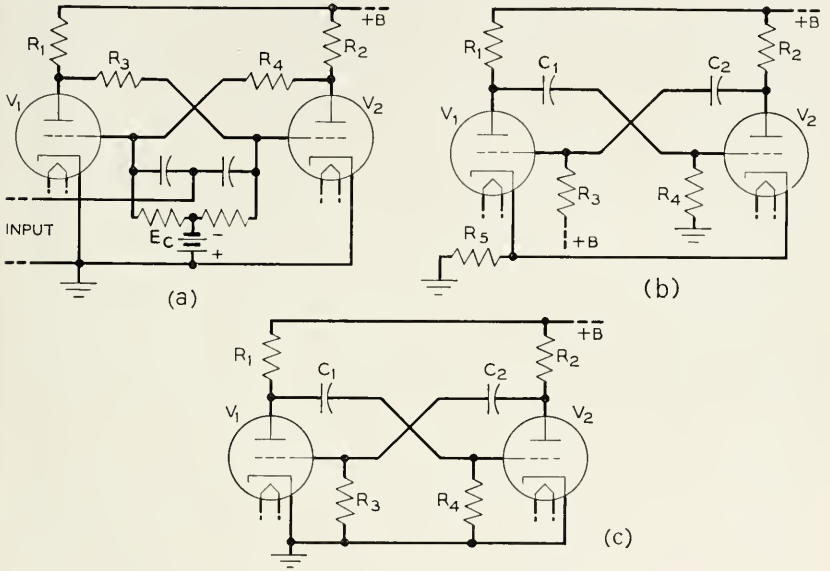


Fig. 47.—Basic Multivibrator Circuit Forms.

resistances is of a rectangular form whose duration is determined by the time interval which exists between the two applied excitation pulses.

A basic modification of the fundamental trigger circuit which has been found most useful in radar sweep circuit is given in Fig. 47b. Here the original circuit form has been modified so as to permit a complete cycle of operation upon excitation by a single actuating pulse. The duration of the cycle is here internally controlled by the arrangement and value of the circuit elements. This form of multivibrator is characterized by having one stable equilibrium condition and is known as a "one shot" type.

The chronological order of operation of this circuit type may be considered as follows: V_1 is normally conducting heavily because of the large positive grid potential impressed upon it by the plate supply battery and the connec-

tion through R_3. The plate current of V_1 flowing through the common cathode resistor R_5 results in a large effective bias applied to V_2 which continues to maintain V_2 in a cut-off condition. If a negative pulse of relatively short duration is impressed upon the grid of V_1 this tube will be driven toward cut-off with an attendant increase in the plate potential of V_1. This positive increase in voltage will be impressed upon the grid of V_2 causing V_2 to conduct plate current. The resulting decrease in the potential at the plate of V_2 further decreases the grid potential of V_1 through the coupling condenser C_2. This action progresses until V_1 is driven beyond plate current cut-off and V_2 is conducting. This condition remains as long as the discharge of the condenser C_2 through R_3 will maintain the grid of V_1 at a net negative potential. When the condenser C_2 has discharged sufficiently to allow the grid of V_1 to increase above the cut-off value, V_1 will again conduct and the resultant action will reduce and eventually cut off the plate current of V_2. The duration of the cycle of operation is here shown to be dependent on the time constant of the circuit R_3 C_2 and may accordingly be controlled as desired by proper selection of these elements. The return of the grid of V_1 to a very high positive voltage point in the circuit has a definite advantage which may be considered as follows: A variation of the grid voltage of V_1 required to cut off the plate current will influence the time duration of the cycle of operation. Here the time rate of change of the grid voltage has been made extremely large by the choice of the return to the high-voltage supply. Thus, an order of magnitude increase in the duration stability of the circuit is achieved.

A further modification of the trigger circuit furnishes the third general type of multivibrator employed in the radar receiver field. This circuit form, called the "free-running" type, has the property of presenting two unstable limiting conditions and accordingly will produce sustained oscillations of a nonsinusoidal form. Figure 47c illustrates this circuit arrangement. The essential circuit change over that given in Fig. 47b, is seen to be the elimination of the stable equilibrium condition of V_1 by the absence of a positive potential on the grid of V_1.

In the free-running type of multivibrator shown, the duration of operation of a particular tube is related to the time of discharge of the coupling condenser and the grid resistance associated with the tube. If a different time constant is chosen for each tube circuit, an unsymmetrical wave form, i.e.—a pulse-to-no-pulse interval ratio other than one, can be produced. In general, the free-running multivibrator is seldom used in this basic form because of the limited repetition-rate stability of this circuit. It is customary, however, to trigger this free-running type of multivibrator with short-duration pulses having a slightly higher repetition-rate than that determined by the multivibrator circuit constants. In this manner the repetition-rate may be

externally controlled as desired. It is also possible to synchronize this particular form of circuit at a submultiple of the externally available trigger repetition frequency.

Pentodes and other now available multi-element vacuum tubes, where the multivibrator interstage coupling involves additional control elements, are commonly employed in the modern radar receiver. Wave forms other than the basic rectangular pulse forms appearing at the plate terminals of the multivibrator circuit are available at various other points in the circuit and are often employed in specific applications.

One other basic form of pulse producing electronic circuits is known as the "blocking-oscillator" type: two typical examples of which are illustrated in Fig. 48. Here the positive feedback of energy required to produce the multivibrator characteristic is realized through the use of a single vacuum

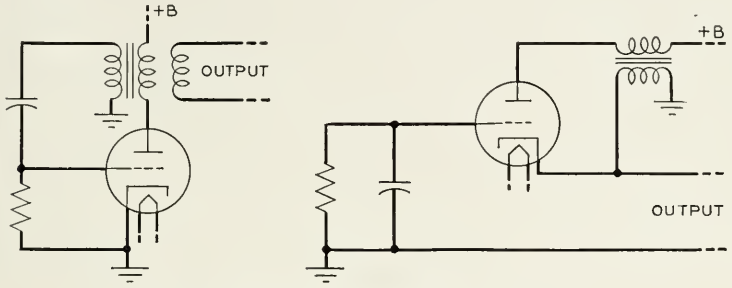


Fig. 48.—Typical Blocking-Oscillator Circuit Forms.

tube and a transformer feedback circuit. This form may be described as an oscillatory vacuum tube circuit where the grid circuit is so arranged to be driven negative after one or more cycles of operation. This results in an intermittent oscillation and the production of nonsinusoidal wave forms similar to those produced by the general multivibrator circuits previously described. The basic advantage of the blocking oscillator circuit form is one of economy of vacuum tubes and attendant power supply reduction.

Typical Timing Wave Circuits

The practical military equipment requirements of World War II with the emphasis on compactness and low-power consumption has resulted in the development of a myriad of specialized circuits which reflect the ingenuity of the electronic circuit designer and the basic flexibility of the modern vacuum tube. In general, however, these circuit developments are quite similar operationally to the basic forms here described.

Figure 49 illustrates a typical circuit arrangement of the sweep timing portion of a PPI indicator as employed in a naval search radar equipment.

In this particular radar system the transmitting magnetron is pulsed from a free-running modulator and, therefore, the controlling timing reference pulse for sweep purposes must be obtained from the modulator circuit. In many other military radar systems it has proven desirable to time both the transmitter modulator and the receiver sweep circuits from a common controllable repetition frequency source. As shown in Fig. 49, a positive synchronizing pulse as obtained from the transmitting modulator is delivered to the radar receiver for range timing reference and here applied to the "clipper" portion of the timing circuit. It was considered here desirable to clip or limit the timing pulse to gain freedom from timing instability, due to possible amplitude variations and to eliminate any possible negative excursions of the timing pulse which might cause faulty operation of the following multivibrator circuit. The multivibrator shown is a modified form of the "one-

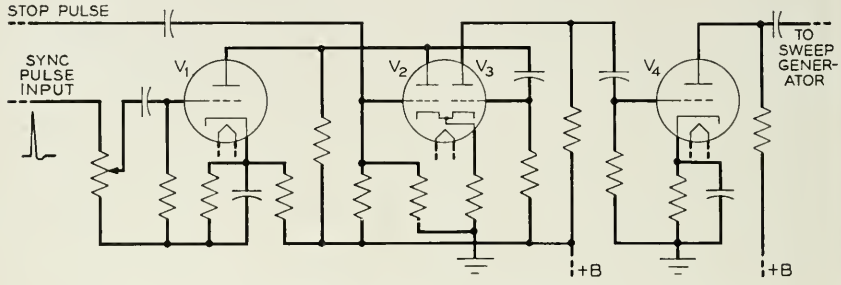


Fig. 49.—Radar Sweep Timing Circuit. Simplified schematic diagram.

shot" type described previously. The grid of V_3 is normally maintained at a positive potential through its connection to the positive plate supply source and accordingly V_2 is normally cut off.

Upon application of the positive synchronization pulse to V_1 and the resultant lowering of the plate potential of V_1 the grid of V_3 is driven below cut-off decreasing the voltage drop across the common cathode resistance and causing V_2 to conduct. This condition will be maintained until the coupling condenser has discharged sufficiently to permit V_3 to again conduct. In the circuit here described, however, this controlling time constant has been selected to be somewhat larger than the total period of the sweep rate and the termination of the sweep timing pulse is accomplished by an external stop pulse applied as shown. This stop pulse is developed in the following sweep amplifier circuits not here shown and is controlled directly by the deflection current. Details of this stop-pulse timing and generation is given in a later section.

The output of this timing circuit shown here then is observed to consist of a rectangular pulse whose leading edge is related to the time of the out-

going radar pulse and whose duration has been controlled by the limits of deflection desired on the radar indicator tube. This form of sweep circuit is known as a "start-stop" type and has proven extremely satisfactory as employed in a number of military radar equipments designed during the past war period.

2.53 *The Sweep Wave Form Generator*

The sweep wave form generator is required to generate the specific voltage or current time functions required to properly deflect the electron beam of the cathode-ray display device. The timing of the interval of this sweep wave form is provided by the timing or synchronizing circuits just described.

In general, it has been required that the range sweep wave form amplitude be essentially a linear function of time over the range interval under observation. During the latter portion of the war, certain airborne applications of radar did require that a specific nonlinear wave form be employed, but the commonly employed displays (A, B, C, and PPI) are usually operated with linear range deflection sweep circuits.

The basic method of obtaining a sweep voltage wave form which increases with time is illustrated in Fig. 50a. In this circuit V_1 is normally operating at little or no bias and, therefore, due to the large voltage drop across the plate resistor R, the plate potential of V_1 is considerably lower than the plate supply voltage B. If a negative rectangular pulse is applied to the grid of V_1 the tube will be abruptly driven to cut-off and, due to the current flow through the condenser C, the plate potential will rise exponentially as indicated to eventually assume the value of the supply voltage B. At the time of end of the negative driving pulse, V_1 will again conduct and the potential at the plate of V_1 will be abruptly reduced as shown.

There are several methods employed in radar sweep circuits to improve the linearity versus time of the fundamental exponential sweep wave form. The first of these takes advantage of the fact that the initial rise of the exponential wave form in the limit is a linear function of time. By using only a small portion of the wave form shown and supplying later amplification to produce the desired deflection, a simple improvement results. This form of linear sweep generation represents the original and by far the most common of the types employed in military radar systems during the past war.

Figure 50b illustrates a method of improving the linearity of the sweep wave form whereby the exponential wave form generated by the basic condenser charging operation is modified through the use of feedback. As shown here the asymptotic value of the exponential charging voltage has been increased by a factor of $(\mu + 1)$ and the effective time constant of the charging circuit has likewise been increased by the same factor. The use of an amplifier in the feedback circuit having an effective gain of 50 would

result in an improvement in linearity comparable with the circuit of Fig. 50a where the plate supply voltage was increased by the same factor.

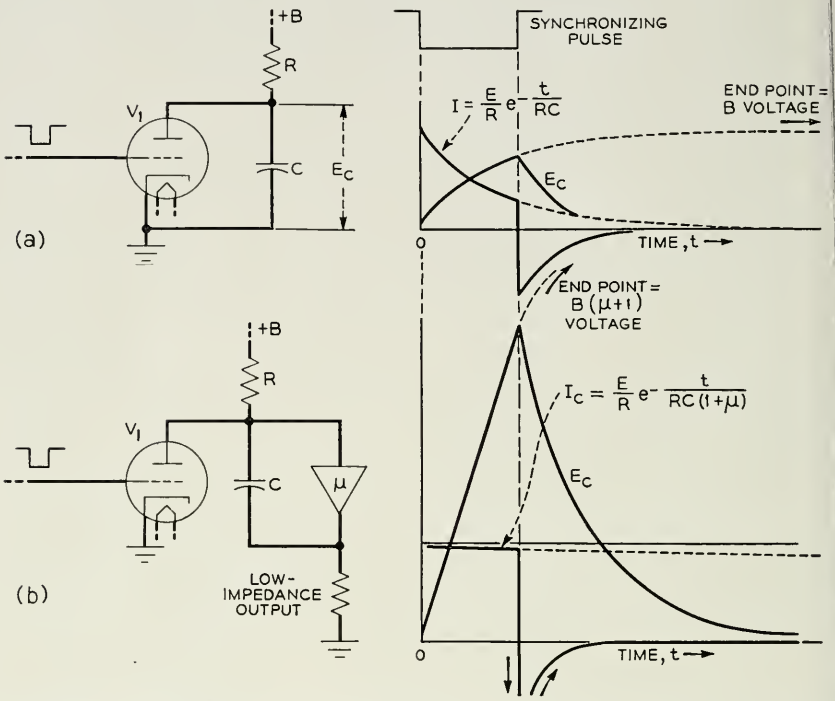


Fig. 50.—Radar sweep generation circuits—basic form and modified by negative feedback to improve linearity.

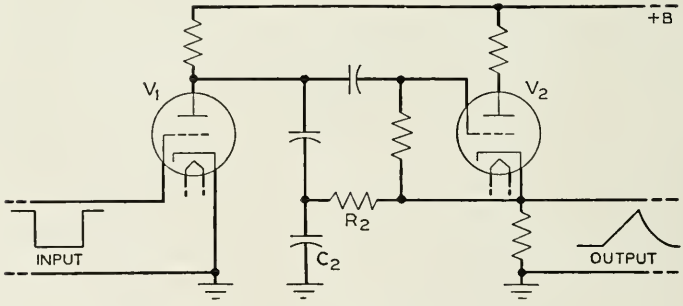


Fig. 51.—Circuit employed to improve linearity of sweep wave form by integration method.

A further method of improving the linearity of the generated sweep wave form is illustrated in Fig. 51 where an additional correction voltage is superimposed on the exponential sweep wave form. This correction voltage is

derived by integration of the sweep wave form as impressed upon the elements R_2, C_2 and the voltage appearing across C_2 is effectively superimposed upon the output wave form. As employed on an airborne bombing radar equipment, a circuit similar to that shown in Fig. 51, was employed where a residual nonlinearity of less than 0.5% was achieved and maintained under severe military operating conditions.

In certain instances it is desirable to generate a sweep wave form which has a specific nonlinear time characteristic. An illustration of one such case as applied to airborne radar is given in Fig. 52. Here the airborne radar display was required to present a nondistorted ground plan which in

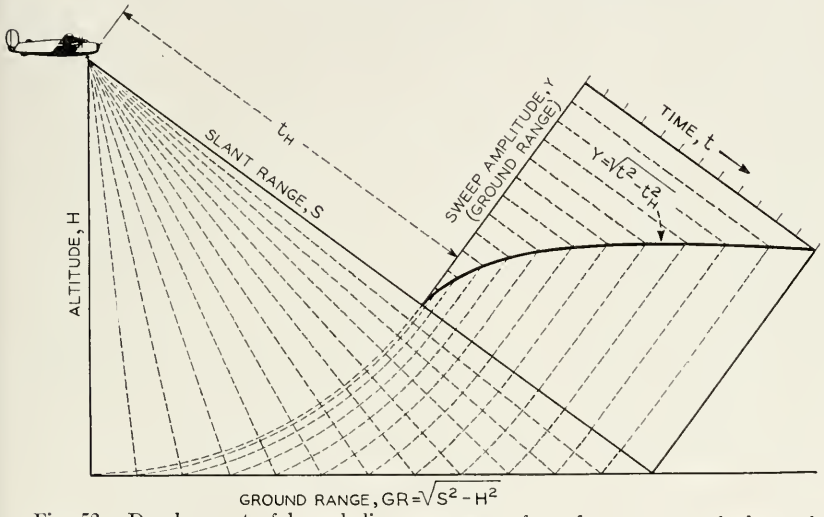


Fig. 52.—Development of hyperbolic sweep wave form for true ground plan radar presentation.

turn required that the range sweep wave form be of a hyperbolic form. The start of the display sweep must be delayed in time corresponding to the time of propagation and return of the radar pulse between the aircraft and the ground. This delay may be produced by the use of a multivibrator of a convenient form, actuated by a pulse coincident with the outgoing radar pulse, and where the duration of the multivibrator pulse is controlled either manually or automatically by reference to the aircraft's altimeter.

The hyperbolic sweep wave form illustrated may be approximated mathematically as the sum of a linear and a series of exponential terms. In this particular application, it was found sufficient to consider a linear and two additional exponential terms only to satisfy these specific requirements. Figure 53 indicates the method employed to generate this specific wave form. As indicated, the desired theoretical hyperbolic sweep function has

an infinite starting slope which cannot be provided with the practical limitations of frequency band width and power available so that here the actual delay used was chosen as 0.9 H, resulting in evident but acceptable distortion in the display in the area directly beneath the aircraft. Figure 53b indicates the fundamental circuit method of generating this wave form. The linear term is generated across the capacitance C_0 and the series current flowing through the additional elements R_1, C_1 and R_2, C_2 supplies the two

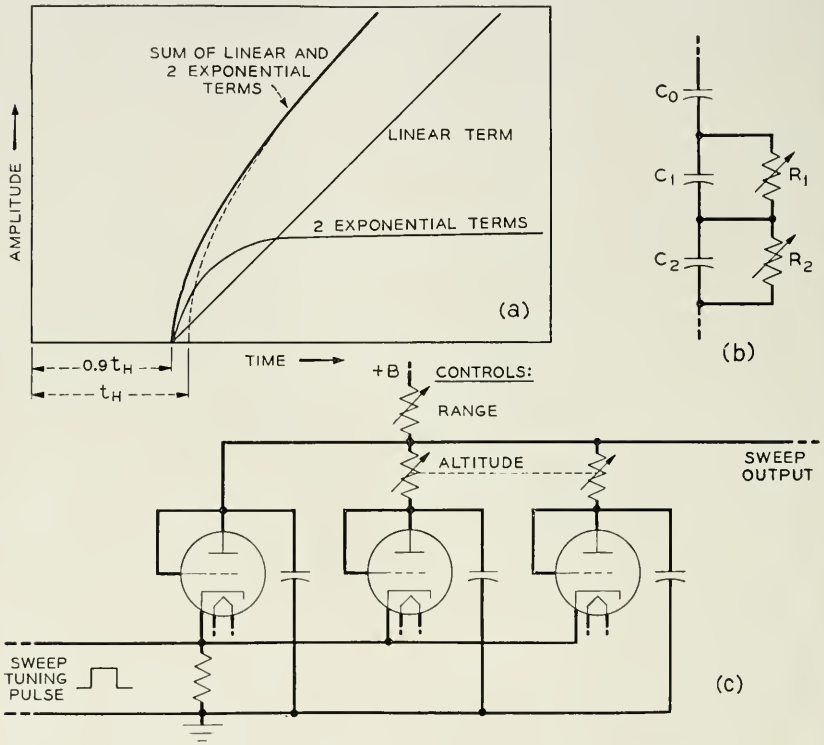


Fig. 53.—Hyperbolic sweep wave form generation—simplified schematic diagram.

additional exponential voltage wave forms. The resistances R_1 and R_2 are required to be variable, their value being determined by the altitude of the aircraft. The practical form of the circuit employed is outlined in Fig. 53c, which includes the additional resistor required to enable modifying the rate of rise of the sweep wave form in accordance with the selected interval of ranges to be displayed.

2.54 *The Sweep Amplifier*

The remaining portion of the radar sweep circuit is concerned with the amplification of the properly timed and generated sweep wave forms to

assure adequate deflection voltage or current for display purposes. The sweep amplifier for range deflection purposes is essentially a specialized form of video amplifier which must be capable of wide band transmission to adequately reproduce the short time sweep wave forms and whose output characteristics are such as to properly supply the high voltage or current signals as required by the radar display system. Two general sweep amplifier design problems are presented for the two basic radar indicator types. The

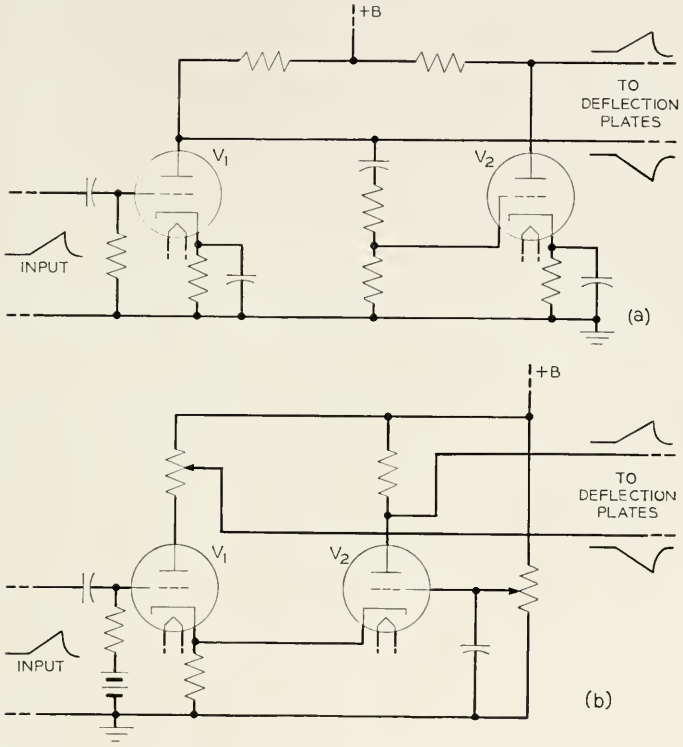


Fig. 54.—Range sweep amplifier circuit schematics for electrostatic-type radar displays.

electrostatic type cathode-ray tube generally requires a balanced to ground deflection signal of moderately high amplitude while the magnetic type cathode-ray tube requires a large deflection current for its operation.

Figure 54a illustrates a simplified schematic of a range sweep deflection amplifier to be employed with an electrostatic type-A radar display. Here the previously generated sweep wave form is impressed upon the grid of V_1 and after amplification a portion of the signal of opposite polarity and of amplitude comparable with the input signal at the grid of V_1 is applied to the grid of V_2. The plate circuit of each tube is connected directly to the deflection plate of the electrostatic cathode-ray tube. In this instance,

the average potential of the horizontal plates of the indicator is maintained at a value determined by the d-c plate potentials and, as indicated previously, this same potential should be applied to the second anode of the cathode-ray tube to avoid defocussing effects. Another variation of a phase inverter amplifier which is commonly employed in radar sweep circuits is illustrated schematically in Fig. 54b. In this instance, a common cathode impedance is employed to accomplish similar excitation of the balanced output tubes. If one grid is excited the plate current flow of this

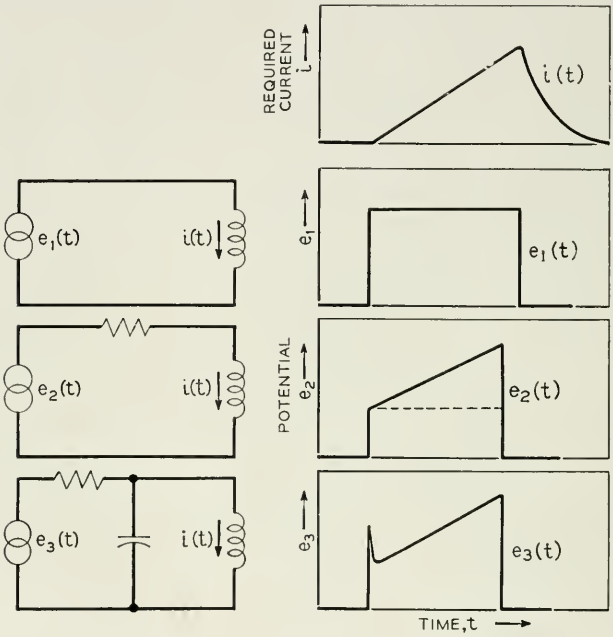


Fig. 55.—Voltage-current-time relationships for magnetic deflection structures.

tube through the cathode resistance serves to excite the second tube and a balanced output sweep signal voltage will result. The values of the plate resistors in this form of circuit are of unequal values and must be adjusted to produce the desired balanced output. The additional control illustrated in the grid circuit of V_2 may be employed to serve as a d-c positioning control.

The sweep amplifier design considerations involved for the magnetic deflection type of cathode-ray tube radar indicator are somewhat more involved, due primarily to the character of the amplifier load impedance. In the case of magnetic deflection the final flux density, and accordingly the sweep current through the deflection coil, is required to be proportional to

the deflection time function desired. If a linear deflection function is assumed, as shown in Fig. 55, producing a linear sweep, the necessary form of the applied voltage wave will vary depending on the inductance, the resistance and the parasitic capacitance of the coil circuit. These conditions are illustrated in Fig. 55. It is entirely possible to generate sweep voltage functions of the forms indicated here for application to a linear amplifier and deflection coil circuits and in fact such an approach was employed in early military radar designs of World War II. It has, however, proven more convenient to employ negative feedback amplifiers whereby the deflection coil current output is maintained proportional to the applied voltage at the input of the sweep amplifier. In this manner, a sweep generator voltage wave form can be employed which has the characteristics desired of the final deflection.

A simplified schematic of a feedback sweep amplifier to be employed in connection with a magnetic deflection radar indicator is shown in Fig. 56. In this example the impressed sweep wave form voltage having the essential characteristics of the desired deflection time function is impressed upon the grid of V_1. This sweep form is amplified and the deflection coil current of the output stage which flows through the 80-ohm cathode resistance common to the first and third stages produces a voltage drop proportional to this current which is effectively applied between the cathode and grid of the first stage thus completing the negative feedback loop. If sufficient forward gain and adequate feedback is provided, the deflection coil current can be made to assume the essential characteristics of the original impressed voltage sweep wave form. It should be observed that the grid of the third stage is biased negatively beyond plate current cut-off to insure that the deflection coil current has an initial value of zero before the start of the sweep. If this condition is not observed, the zero range point on the indicator will be a function of the d-c current of the output stage and in the case of a PPI form of display, the zero range region will assume the characteristics of an open circle and map distortion at the short ranges will result. In this amplifier circuit, application of the sweep signal to the grid of V_3 will not result in deflection current flow until the tube is driven above cut-off. During this time the feedback is not effective and the over-all gain of the amplifier is at its maximum value. Due to the inductive characteristics of the amplifier load impedance, the initial rise in current will be delayed slightly with respect to the applied voltage and accordingly a further delay of the feedback voltage is introduced by the use of a time constant in the common feedback network. The result is a delaying of the applied feedback voltage with a corresponding period of maximum gain of the amplifier which tends to produce a sharp increase in deflection coil current at the time of the

start of the sweep. After this short interval of time, the feedback becomes effective and the output current and input voltage correspondence obtains.

It is essential in this type of circuit that the feedback be determined entirely by the deflection coil current if optimum operation is to be obtained. It should be observed that this condition requires that the screen current which also normally flows through the feedback impedance does not con-

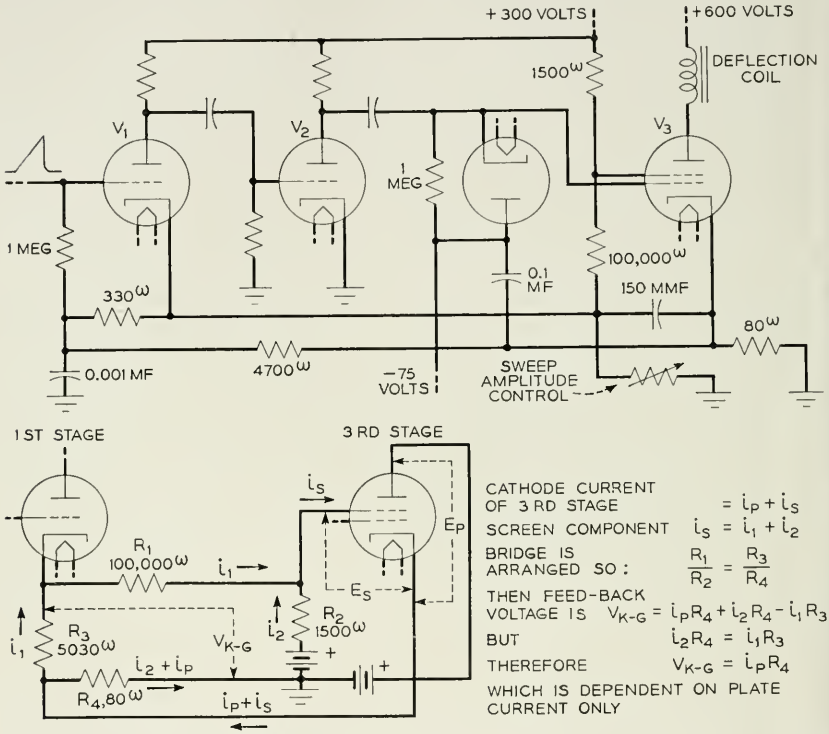


Fig. 56.—Radar range sweep amplifier employing negative feedback and screen grid bridge circuit—simplified schematic.

tribute to the net feedback. Figure 56 illustrates the bridge circuit which has been devised to accomplish this. A voltage from the screen of the third stage is directly applied to the cathode of the first stage, this voltage being equal in magnitude and of opposite polarity to that which appears across the feedback impedance due to the third stage screen current flow. This bridge circuit operation is independent of the absolute screen voltage value.

To insure identical starting potential conditions regardless of the duration of the range sweeps in use, d-c restoration is employed in the grid circuit of the last stage. The action here is similar to the operation described previously in connection with video amplifier design. The delay inherent in

the magnetic deflection circuits must be carefully considered in the overall radar receiver design, if the display is required to reproduce short-range information. In such cases, it is customary to insert delay networks in the video channel introducing a delay to the received signal equal to that present in the indicator deflection system, or to "pre-pulse" the deflection circuits prior to the time of the outgoing radar pulse.

It is desirable from a power consumption and display appearance standpoint to limit the range deflection of the radar indicator only to that amplitude required to adequately fulfull the display requirements. A method commonly employed is indicated in Fig. 57. Here a measure of the current flow through the deflection coil, and accordingly the amplitude of the deflection of the cathode-ray tube beam upon the screen, is obtained from the feedback voltage of the sweep amplifier. This voltage is impressed upon a

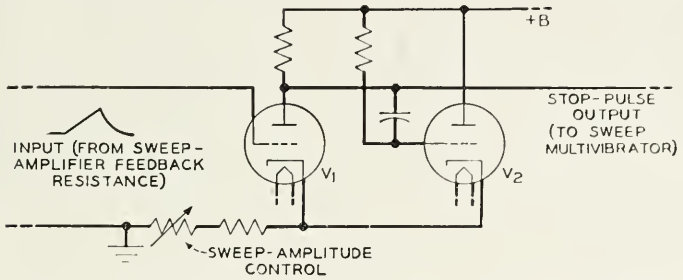


Fig. 57.—Range sweep stop pulser circuit for limiting sweep deflection.

"sweep-stop" pulser which upon rising to a preselected value corresponding to the desired sweep amplitude is caused to trigger this circuit. The output pulse of this circuit is then employed to operate the sweep limiter portion of the sweep timing multivibrator previously described, thus terminating the sweep timing pulse proper.

2.6 *Circuits for Radar Range and Bearing Measurement*

In this review of radar receiver design principles only the presentation of the received radar signal in a form convenient to the observer has been considered. To fully utilize the complete radar information available, determination of the complete coordinates is necessary with an exactness which is determined by the specific use of the data and by the capabilities of the radar system itself. This section will be devoted to a review of the methods employed to generate electronic markers necessary for the determination of range and azimuth and elevation angles. These markers include both the fixed variety, whereby the approximate coordinates of a radar target can be determined by inspection, and steerable markers by

which means precise coordinate data necessary for most military applications are determinable. As indicated previously, the optical filters commonly employed over the screen of the radar indicator often serve as a medium for display of range, azimuth or elevation coordinate markings: however, these methods are seldom completely satisfactory in military fire-control radar systems because of errors introduced by the ever-present size or position variations in the electronic display field. Their use has been strictly limited to search or reconnaisance radar systems.

2.61 *Electronic Bearing Marker Circuits*

The bearing marker methods reviewed here are applicable generally to both azimuth and elevation angle determination. A method of azimuth or elevation bearing determination which can be associated with a lobing antenna system and an A-type indicator has been mentioned previously. This method remains an extremely precise system which has the desirable advantage of simplicity. During the latter part of the war, automatic tracking was applied to this method where the actual comparison of the lobes of the selected target signals was carried out electronically and the resultant antenna steering information utilized as the final bearing data. In a strict sense, however, only an indication of error in antenna training is observable to the operator on the radar equipment proper. The exact bearing data must be obtained from a measurement of the position of the antenna itself.

In the case of the continuous scanning systems employing B, C, or PPI presentations, it is common practice to provide a steerable electronic marker which can be superimposed upon the display field and by which means relatively exact azimuth and elevation angles can be determined by target and marker coincidence. This electronic marker method has the advantage that it is subject to the same size and position display field distortion influences as the received pulse signal, thus eliminating this source of error.

A circuit arrangement which has been employed in connection with a naval vessel radar search system to brighten a selected and variable range trace of the PPI indicator to serve as an electronic azimuth marker is given in Fig. 58. In this example the rotating antenna structure is equipped with a small permanent magnet pole piece whose cyclic excursions past a sealed magnetic reed relay cause a pair of contacts to close indicating coincidence. The relay structure is likewise mounted on a ring which can be rotated with respect to the scanning axis of the antenna. The relative bearing of a target is thus determinable by a knowledge of the angular position of the relay capsule with respect to the lubber line of the vessel. The circuit of Fig. 58 produces one brightened range trace for each revolution of the antenna upon closure of the bearing marker relay contacts and is ar-

ranged to be unaffected by any subsequent chatter or false switch closures. The pedestal generator which includes vacuum tubes V_1 and V_2 which are normally operated below plate current cutoff produces upon closure of the bearing marker switch contact a rectangular negative pulse having a duration of 10,000 microseconds. This pulse operation is independent of additional chatter effects following the initial contact closure. The following single-cycle bearing mark multivibrator is normally held inoperative by the bias voltage developed on the cathode of V_3. The grid of V_3 is continuously excited with the range sweep start pulses of an amplitude insufficient to actuate this multivibrator circuit. The presence of the 10,000-microsecond pedestal is sufficient, however, to allow the following range sweep

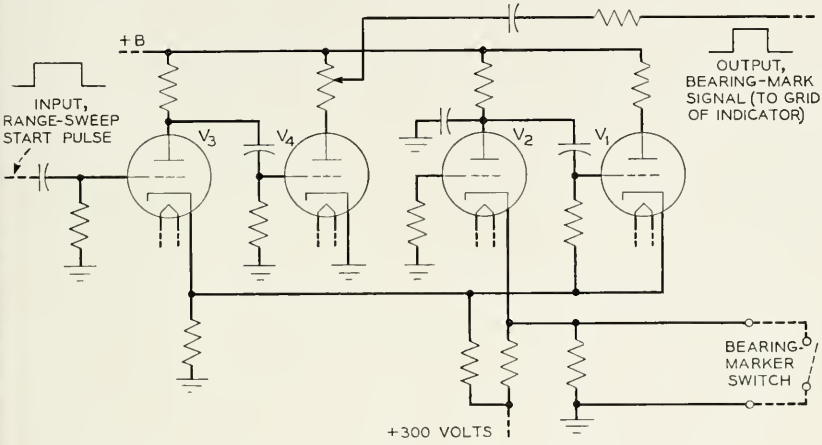


Fig. 58.—Electronic azimuth bearing marker circuit—simplified schematic.

start pulse to operate the multivibrator. The output of this circuit is then a 550-microsecond pulse which represents a time slightly longer than the maximum range to be displayed (60,000 yards) but shorter than the period of the sweep repetition rate. This positive 550-microsecond pulse is applied to the modulating grid of the PPI indicator tube through an adjustable trace brightness control element.

Another convenient azimuth display method which has been extensively employed in naval and airborne radar systems involves the use of "true North" presentations. Here the PPI azimuth indication is presented in terms of a compass reference, the actual instantaneous position of the range trace on the screen representing the compass direction of the antenna beam. In the indicator previously illustrated in Fig. 44 the compass information is introduced by means of a differential synchro inserted in the antenna-indicator synchronizing connections whose angular displacement is pro-

portional to the instantaneous heading of the aircraft. In the SL radar indicator shown in Fig. 61 the compass information is introduced by means of a mechanical differential rotation of the indicator deflection coil proportional to the angular position of the compass repeater mechanism.

2.62 *Range Marker Circuits—Fixed Range Markers*

A simplified schematic diagram of a convenient fixed range mark generator circuit which has been extensively employed on airborne radar search systems is given in Fig. 59. Here the radar system requires 1- and 5-statute mile fixed range markers. The sweep start multivibrator pulse is applied to the grid of V_1 as shown. In the absence of a signal, this tube

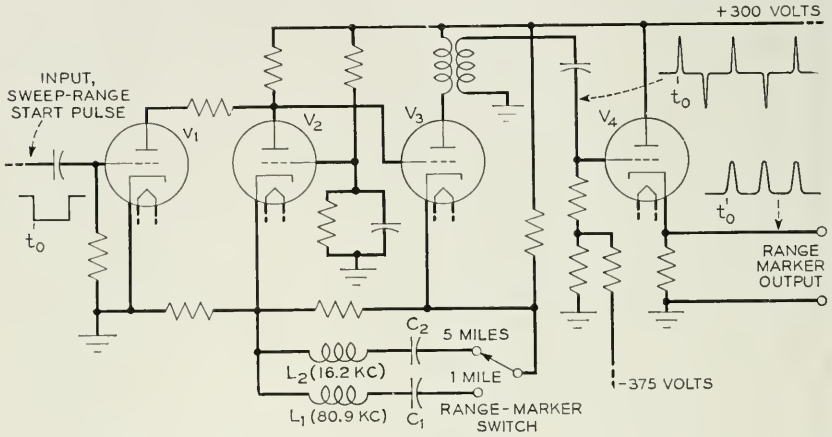


Fig. 59.—Fixed range marker circuit—simplified schematic.

operates at effectively zero bias, and because of the large plate current flow reduces the effective plate potential of V_2 and the grid potential of V_3 to a low value. Since the cathode of V_3 is subject to a large positive potential, this tube is cut off and the oscillatory circuit shown is inoperative. Upon application of the negative start pulse, V_1 is cut off for the duration of this pulse and the attendant rise in the V_3 grid potential permits the oscillator circuit to function. The series resonant elements $L_1 C_1$ and $L_2 C_2$ determine the frequency of oscillation by providing a high value of positive feedback at the series resonant frequency. The output of V_3 which consists of approximate sinusoidal pulses is differentiated by means of the air-core transformer shown. The differentiated pulses are then applied to a cathode follower amplifier stage biased to cut off where the output is limited to the desired positive fixed range mark pulses for indicator display. By careful choice of circuit elements and equipment arrangement, this simple circuit form has

produced an entirely satisfactory range marker signal for radar search systems.

Variable Range Marker Circuits

Variable range marker circuits are employed where more precise range information is required for missile control applications of radar. Here the observer may position the electronic range mark to obtain coincidence with the selected target, and from an associated calibration of this positioning control, determine the range coordinate. For search or reconnaissance purposes it is often desirable to determine range with somewhat more accuracy than is afforded by the display of fixed markers, and for this purpose,

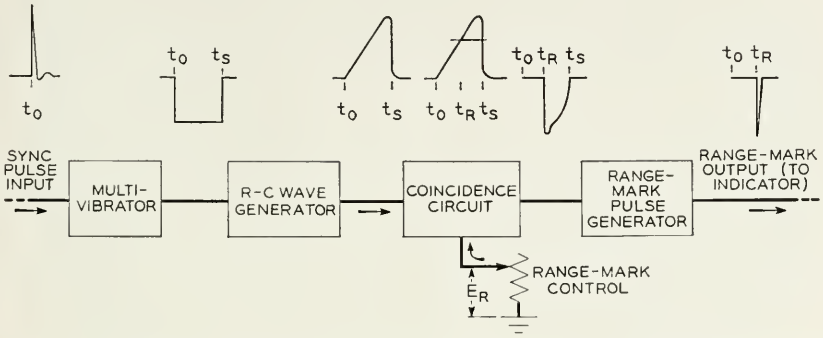


Fig. 60.—Variable range marker circuit of moderate precision—simplified schematic.

several designs of moderate precision variable range marker circuits have been developed and employed during the past war period.

Figure 60 illustrates the circuit operation of a variable range mark generator of moderate precision. This circuit operation depends on a pulse obtained from the transmitting modulator to serve as the zero time or range reference. This is applied to a single-cycle multivibrator which produces a rectangular pulse whose leading edge is coincident with the time of the synchronizing pulse and whose duration is somewhat greater than the maximum range measurement required. A saw-toothed voltage wave form is generated in the following RC wave generator by means similar to the sweep wave form generation described in a previous section of this paper. The coincidence circuit which follows consists of a vacuum tube biased below cut-off whose exact cut-off bias is determined by the range mark potentiometer setting. This coincidence circuit is thus inoperative until the saw-toothed input signal has reached the value of the cut-off voltage, at which time this circuit functions and produces a sharp decrease in its plate potential. The effective time delay which is here produced with respect

to the time of the synchronizing pulse is observed to be a function of the rate of change of the saw-toothed wave form and the setting of the range control potentiometer which may be calibrated directly in units of range to the target. The following range mark generator differentiates this coincidence circuit output wave form and furnishes the desired amplification. Zero range calibration is here provided by employing a sample of the zero time reference pulse and introducing this voltage into the range control circuit.

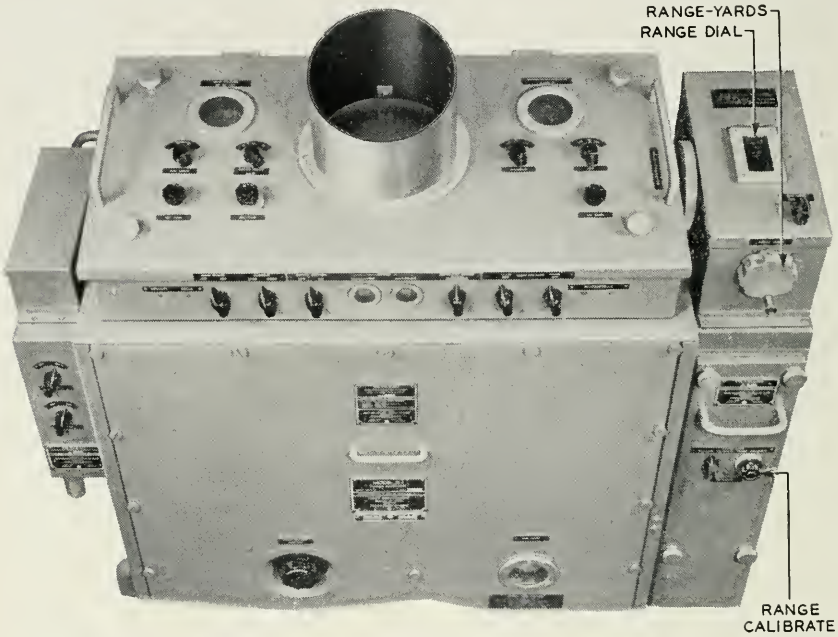


Fig. 61.—Transmitter-receiver-indicator assembly as designed for SL-Naval Search Radar equipment.

Figure 61 illustrates the transmitter-indicator assembly of the SL naval vessel search radar system. This system employs a PPI display with available range sweeps of 5, 25 and 60 nautical miles. The assembly shown to the right of the main unit contains a variable range marker circuit of the type just described. This range mark is positioned by means of the control located toward the top of this unit and its calibration is observable through a window located on the top panel. Here a maximum measuring range interval of 40,000 yards is available. In this application, the RC elements of the wave generator are enclosed within an oven and maintained at a constant temperature by thermostatic means. The accuracy achieved in this example, without recourse to calibration means involving targets at known

range, is ± 200 yards at the maximum range with an accuracy of ± 100 yards for targets within 5 nautical miles.

For more precise determination of range than is afforded in the circuit just described, two methods have been extensively employed. The first method involves the production of a known time delay by actual measurement of the time of propagation of an acoustical wave through a liquid medium. Here the physical length of path is varied to produce the variable delay. The second method involves the phase shifting of a known precise sinusoidal frequency standard which bears a fixed phase relationship to the time of the outgoing radar pulse.

The "liquid delay tank" variable range unit over-all operation may be observed by reference to Fig. 62. The zero time range reference is obtained in the form of a pulse coincident with the outgoing radar pulse. This pulse actuates the one-cycle multivibrator shown to produce a sharp high-ampli-

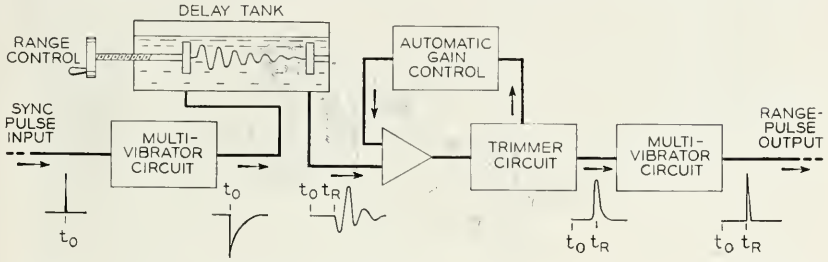


Fig. 62.—Liquid delay tank type of precision variable range unit—block diagram of operation.

tude output pulse, here relatively independent of amplitude and form characteristics of the synchronizing pulse and which is applied directly to the delay tank. This delay tank consists of a suitable container filled with a mixture of iron-free ethylene glycol and water so composed as to produce a zero temperature-velocity coefficient at 135°F, at which temperature the liquid is maintained by thermostatically controlled electrical heaters. In this temperature region the temperature-velocity characteristic is such as to produce a decrease of velocity of 0.1% for a temperature variation of 14°F. Located at one end of this tank is a quartz crystal, approximately $\frac{7}{8}''$ square and 0.040″ in thickness, mounted securely on a brass plate which serves as one electrode and which is immersed in the liquid. A similar crystal element is attached to a lead-screw carriage and located so that the face of this crystal is parallel to the fixed crystal. The distance between the crystal faces can be varied by rotation of the lead screw. The sharp voltage wave applied to the transmitting crystal causes it to oscillate in a damped vibration at its natural frequency for longitudinal waves which in this case

is of the order of 1.4 megacycles. The mounting plate and surrounding liquid serves to highly dampen this oscillation. A short vibrational wave train is projected through the liquid toward the receiving crystal. The amplitude of this disturbance is only slightly attenuated by viscous dissipation for the maximum path length here employed. The large area of the crystal relative to the wave length results in a highly directive radiation and is reflected in a parallelism requirement for the crystal faces of the order of .01″. The voltage developed across the receiving crystal upon application of this delayed acoustical wave consists of a main response followed by minor disturbances due to re-reflections between the crystals.

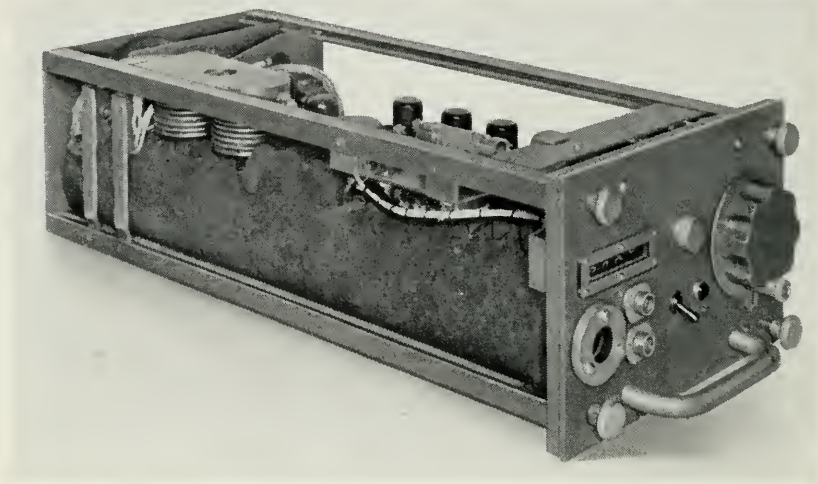


Fig. 63.—Liquid delay tank type of precision variable range unit.

The following amplifier shown in Fig. 62 is required to increase the .005-volt received signal to approximately 20 volts. This gain supplied is controllable by means of an automatic gain control circuit so as to provide a relatively constant amplitude of the first response signal. The following trimmer circuit consists of a pentode operating below cutoff such that a signal of at least 20 volts is required for plate current flow. Since the AGC circuit operates to adjust the gain of the amplifier to this condition, only the first and highest response peak is transmitted to the final range pulse multivibrator circuit where a sharp narrow rectangular pulse is produced to be employed in the following indicator circuit.

Figure 63 is a photograph of the liquid-tank type of variable range unit as developed and manufactured early in the past war and employed extensively in naval fire-control radar systems. This unit includes provision for a

maximum range measurement of 40,000 yards with an accuracy of ± 40 yards at this range under normal field operating conditions.

The use of the liquid range unit is practically restricted to ground and naval vessel application because of its weight and the problems of handling these critical liquids. Another variable range unit development was initiated to meet the same accuracy requirements as above, but to be more suitable for aircraft and other extreme ambient applications. The phase-shifting type of variable range unit whose operation is illustrated in Fig. 64 was the result of this effort.

The input start-stop single-cycle multivibrator circuit produces a rectangular pulse output wave form whose leading edge is coincident with the time of the outgoing radar pulse and whose duration encompasses the maximum range time to be measured, in this example 270 microseconds.

The timing wave generator and associated phase shifting circuit is shown schematically in Fig. 65. The resonant frequency of the oscillatory circuit is 81.955 kc which period represents an equivalent radar range interval of 2000 yards. An initial d-c current of approximately 10 ma is present in the $L_1 C_1$ circuit in the absence of input start signals. Upon application of the negatively poled rectangular start-stop pulse V_1 and V_2 are abruptly driven to cutoff and the energy associated with the magnetic field of L_1 produces local current flow and oscillation at a frequency determined by $L_1 C_1$. The initial circuit conditions here are the same as the zero voltage condition for each cycle of a sustained oscillation and the behavior of the oscillatory system is the same as for the case of sustained oscillation. The absolute average potential of the oscillation is maintained constant regardless of the magnitude of the duty cycle. Positive feedback of the timing wave is included in the V_3 cathode connection in such a manner that uniform amplitude of the timing period throughout the active period results. The purpose of the remaining circuits shown in Fig. 65 consisting of V_3, V_4 and V_5 is to produce four output timing wave voltages whose relative phases differ by 90°. These voltages are to be later combined in such a manner that continuous phase shift of the output timing wave results. Two quadrature voltages are here produced by the use of LR and CR networks so proportioned that $\omega L_2 = \frac{1}{\omega C_2} = R_2$ at 81.955 kc. The desired four timing wave outputs are produced by the use of the phase inverter stages V_4 and V_5.

The method here employed to combine four quadrature voltages to enable continuous relative phase shift of the resultant output is illustrated in Fig. 66. This phase shifter capacitor consists of four quadrant shaped stator sectors which are equal in area and shape and which are mounted parallel to a ring stator as shown. A carefully shaped eccentric dielectric vane rotor

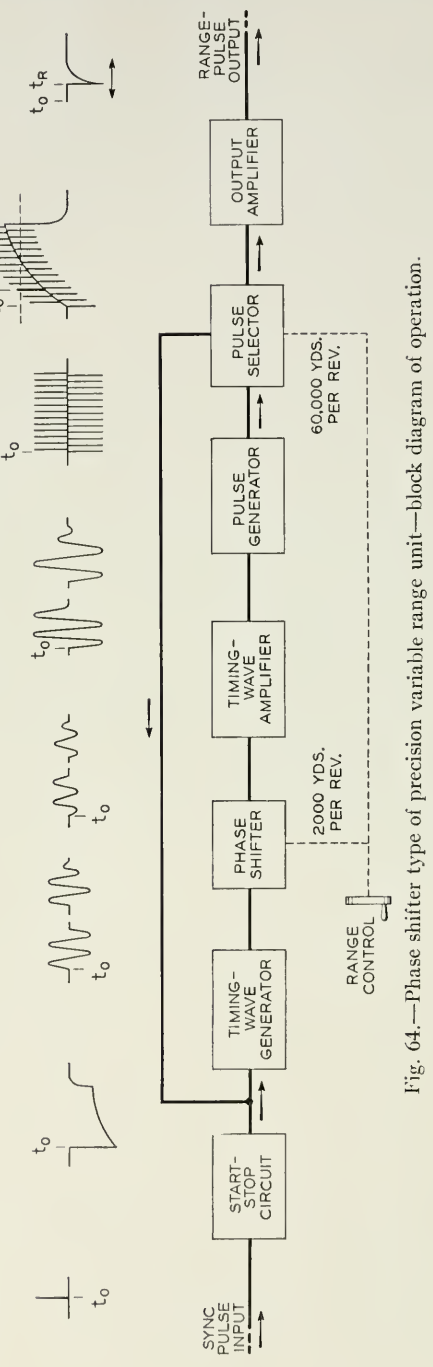


Fig. 64.—Phase shifter type of precision variable range unit—block diagram of operation.

is provided whose rotation between the stator elements affects each quadrant stator capacitance in a like manner. As illustrated the resultant output voltage which appears from the ring stator to ground has a phase shift relative to any applied wave which varies linearly with angular displacement of the condenser shaft.

The function of the following amplifier shown in Fig. 64 is to provide a high-impedance termination for the phase-shifting condenser, and to provide increased amplitude of the timing wave. The pulse generator which follows limits or clips the applied timing wave, and differentiates the re-

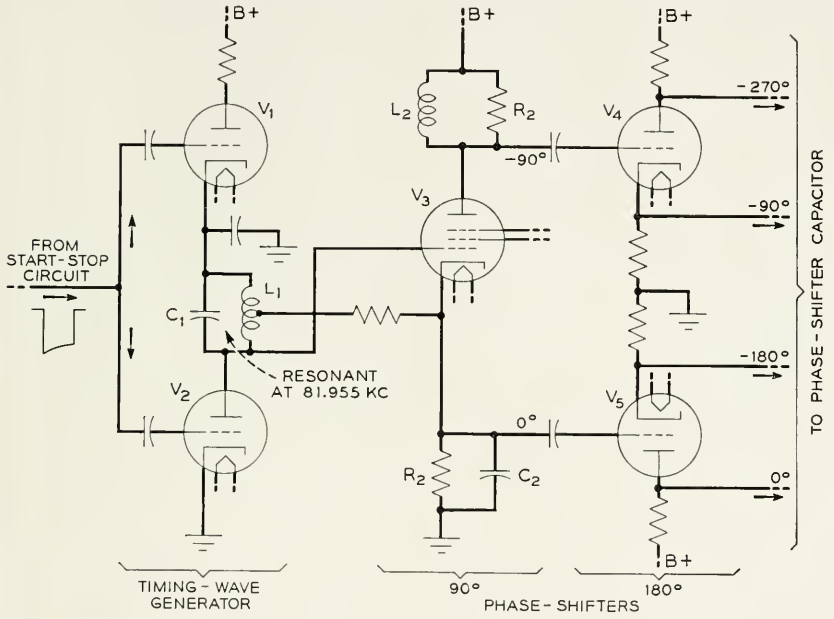


Fig. 65.—Timing wave generator circuit of phase shifter type range unit.

sultant wave form. The output here consists of trains of alternate positive and negative timing pulses.

The pulse selector component shown in Fig. 64 enables obtaining delay intervals greater than 12.2 microseconds the value associated with 360° phase shift of the timing wave. An increasing exponential saw-toothed wave form is generated starting at zero time reference by an RC circuit having a time constant of the order of 800 microseconds. The timing pulses are applied additively with this exponential to the grid of a vacuum tube operating below cutoff, its exact value of bias being determined by the setting of a potentiometer control. At the time that the grid signal amplitude exceeds this critical cut-off bias value, this tube conducts abruptly as

shown and an output pulse is produced whose position on a time scale is determined by the additive phase shift of the timing wave and the setting of the pulse selector potentiometer. By mechanically gearing the potentiometer and phase-shifting condenser, the final rotation of the control shaft will result in an output pulse whose delay will increase uniformly and correspond to 2000 yards per revolution of the control. Further amplification is furnished in the output amplifier shown.

Figure 67 illustrates the final equipment features of a phase-shifting type of variable range unit as developed for naval vessel radar system application. It will be observed that this unit is mechanically interchangeable with the liquid range unit shown in Fig. 63.

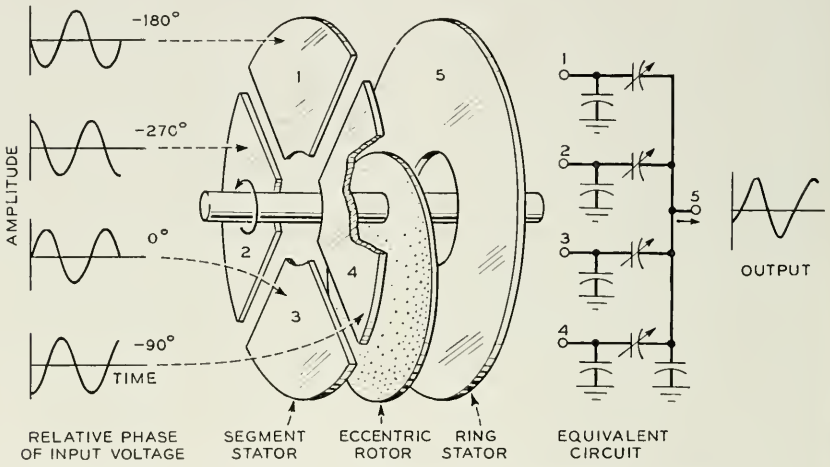


Fig. 66.—Schematic outline of operation of phase shifter condenser.

In this model, the parallel resonant timing wave oscillatory circuit is maintained at a constant temperature by employing an electrically heated oven. Measurements made on this design indicate that the range shift error to be expected for a "warm-up" period of 6 minutes was .003% or 15 yards at 45,000 yards range. After 6 minutes time, thermal equilibrium is reached and the total range error will be less than 20 yards at 45,000 yards range. The unit here illustrated has been universally employed in the majority of naval vessel fire-control radar systems of the past war and these basic circuit principles have served for range measurement in other applications including precision radar bombing.

2.7 *Automatic Frequency Control and Automatic Gain Control*

2.71 *Automatic Frequency Control*

The automatic frequency control (AFC) of the local beating oscillator to

assure correct tuning of the radar receiver has become an extremely important feature of the military radar system and the successful solution of

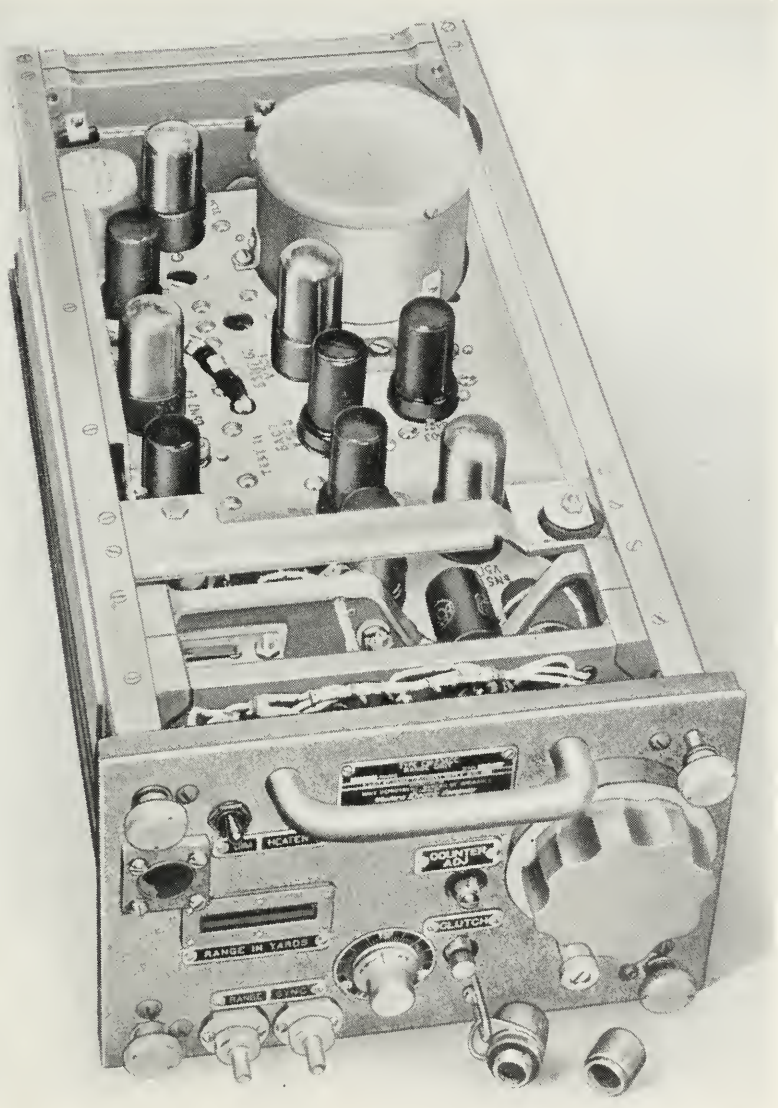


Fig. 67.—Precision variable range unit of the continuous phase shifter type.

this problem has contributed greatly to the practical success of radar during World War II, by assuring consistent optimum system performance under

severe military field conditions. In the case of the usual radio-communication system, some knowledge of character and extent of the information which is being transmitted is available to the receiving location which may serve to evaluate the receiver operating performance, but in the case of the military radar system such reference data is not always available. The usual military operating conditions for radar systems are extremely severe which, in general, tends to degrade the performance. Mistuning of the radar receiver and the attendant reduction of the performance of the system must be immediately evident to the operator even under conditions where no radar signal returns are present.

During the first years of the past war, this tuning problem was recognized and the initial attempts at solution involved the inclusion of receiver tuning indicators to serve as an indication of adjustment. As the radar systems became more complex and with the trend toward the use of higher transmission frequencies the necessity for completely automatic continuous tuning adjustment of the receiver became increasingly evident and the present types of automatic frequency control devices were developed. It has been determined that in the specific case of airborne radar bombing equipment operating at 10,000 mc that the automatic frequency control of the receiver tuning is an absolute necessity, since the radar operator cannot under the normal military operating conditions maintain the system performance in this regard to a small fraction of the optimum.

Functions and Requirements

The basic reference for a radar automatic frequency control system must be the transmitter frequency since it is required that the receiver be properly tuned under the condition where no radar return signals are available. Either the frequency of the transmitter magnetron or the local beating oscillator frequency may be adjusted from an electrical error signal whose characteristics are related to the tuning point. Magnetrons whose frequency was conveniently controllable by remote electrical means were not then available so that the later method has been universally applied in military radar systems developed during the past war period.

It is pertinent to review the nature and extent of the frequency instability of a radar system to derive the requirements to be imposed upon an AFC device. The sources of frequency instability are associated with the transmitter as well as the receiver elements of a radar system. The magnetron frequency is determined in part by the physical dimensions of its oscillatory cavity structure, and as would be expected, ambient temperature and pressure conditions exert a decided influence. For example, a typical thermal coefficient of frequency for a magnetron may be as high as 200 kilocycles per degree Centigrade which over the range of ambient temperatures

to be considered important for military equipment may result in a frequency shift of 20 mc from the time the equipment is turned on until thermal equilibrium is established. The magnetron frequency is extremely sensitive to its terminating load impedance. This termination in a radar equipment is composed of the antenna, the interconnecting RF transmission line, and the duplexing devices. The typical radar antenna system employs rotating joints or connecting devices to enable transfer of RF power to the antenna proper while it is mechanically operated over its scanning cycle. These connections cannot be made to present an entirely uniform impedance over their entire mechanical operating range and thus introduce variable impedance irregularities to the magnetron generator. The frequency of these impedance variations may range from a fraction of a cycle per second to perhaps 60 cps. The input impedance of the antenna proper is dependent on the extent and character of nearby obstructions in the radiation path. The characteristics and form of the radome employed to protect the antenna contribute to the variable impedance characteristics of the antenna and thus influence the magnetron frequency. An additional instability in magnetron operation which is introduced through power supply variations within the modulator and transmitter portion of the system must also be considered in the detailed design of the AFC system.

The receiver itself is responsible for a major contribution to the frequency instability characteristics of the radar system. The local beat oscillator frequency is critically dependent on the physical dimensions of its oscillatory structure and on the supply voltages. The effects of temperature and atmospheric pressure on the frequency of a reflex oscillator of the types previously described is considerable. For example, a thermal coefficient of 0.25 mc per degree Centigrade, typical of the 10,000-mc tubes, will produce a total excursion of perhaps 25 mc over the range of ambients experienced in military equipment. In the case of supply voltage variations, a 5-mc frequency shift will result for a 1% change in anode and repeller potential for a typical 10,000-mc reflex oscillator. Another source of receiver frequency instability is associated with the shift of the IF amplifier frequency selectivity characteristic with tube aging and operating conditions.

If the operating requirements for an AFC system are now reviewed from a consideration of these factors, it will be observed that for a radar system operating at the higher frequencies a total effective frequency change of perhaps as much as 50 mc may be encountered whose rate of change, in general, will be relatively slow and may be classified generally as effects due to "warm up". In addition fast variations of frequency will be present whose rates of frequency change may extend from 1 mc per second per second to 1000 mc per second per second. At the lower radar frequencies

these frequency variations will be somewhat lower with an attendant reduction of the range of operation required of the AFC circuit.

AFC Circuit Design Considerations

To obtain a measure of the basic frequency reference for AFC purposes the direct approach is evident. A sufficiently attenuated sample of the outgoing radar pulse may be obtained from the transmitter and after sepa-

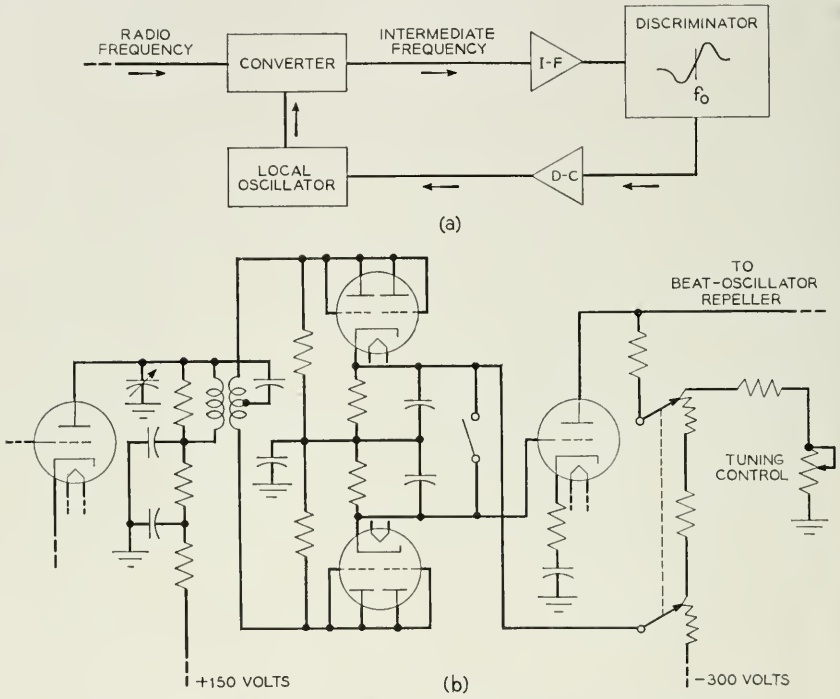


Fig. 68.—Radar AFC system—block diagram and circuit arrangement.

rate conversion but, under the influence of the regular receiver beat oscillator, may be employed as a true sample of the outgoing signal as it exists with the normal receiver IF channel. The separate AFC converter method has been employed in some military radar equipment but has the disadvantage that additional conversion components are required. A second method is more economical of equipment and has been extensively employed during the past war period, but this later type of AFC circuit has limitations which are imposed on it by the character of the IF signal as normally available in a radar receiver. Figure 68a illustrated the essential elements of such an AFC system for a radar receiver. It will be here observed that a

sample of the IF signal after normal conversion and some amplification is applied to a frequency sensitive discriminator circuit and the resulting error signal is employed to readjust the beat oscillator frequency. The outgoing radar pulse is normally attenuated effectively by the TR circuit, and thus the remaining signal available for AFC purposes is due to inherent leakage of the TR elements. As previously indicated, the frequency spectrum of the output "spike" of the TR device extends over a wide frequency range, due primarily to the small finite delay in the breakdown of the TR tube. The energy frequency distribution characteristic of this spike is to a large degree independent of the magnetron frequency and, therefore, must be considered as an undesirable masking signal and accordingly reduced to a noninterfering level. As previously indicated this is usually accomplished by disabling one or more of the IF amplifier input stages for a short time interval coincident with the outgoing radar pulse.

With a signal available which is related to the frequency of the outgoing pulse, the remainder of the AFC design is concerned with the utilization of this information to accomplish the automatic tuning of the radar receiver. To determine the frequency gain characteristic of the discriminator circuit it is pertinent to examine the frequency repeller potential relationship of the local beat oscillator. This relationship for a 2K25-type reflex oscillator operating at 10,000 mc shown in Fig. 19 is approximately 2 mc/volt and is representative of the tubes of this type. This quantity provides an indication of the "loop gain" required for a satisfactory AFC circuit.

Typical AFC Circuit Designs

Figure 68b illustrates the essential elements of a radar AFC discriminator and amplifier circuit. This consists of an input circuit which is required to furnish the means for frequency measurement, rectifier elements to convert this frequency deviation information to a proportional voltage error signal, followed by an amplifier to increase the amplitude of this signal to the required level to adequately control the frequency of the local beat oscillator.

The operation of the discriminator input circuit may be observed by reference to the vector diagram of Fig. 69a. The input circuit, essentially a double-tuned transformer having a low value of mutual inductance, serves to couple the AFC rectifier to the preceding IF amplifier tube. The resonance frequency of both primary and secondary circuits is maintained at the desired midband IF tuning point, in this example, 60 mc. The output of the balanced secondary winding of this input network is applied to a balanced rectifier shown in the vector diagram as E_1 and E_1'. In addition a portion of the IF signal voltage which appears across the primary winding is also applied to each element of the balanced rectifier. At resonance, the primary and secondary voltages assume a quadrature relationship as indi-

cated. The vector relationships for frequencies above and below this resonance, as shown, result in an amplitude change across the rectifier circuit. Figure 69b illustrates a typical rectified voltage versus frequency characteristic of such an array. The location of the actual crossover zero voltage point is determined only by the resonance of the secondary circuit over the limited range under consideration here. The primary resonance contributes essentially only to the symmetry of the voltage output versus frequency characteristic. The introduction of the time constant elements in the detector output circuit integrates the pulse output and are chosen with due regard

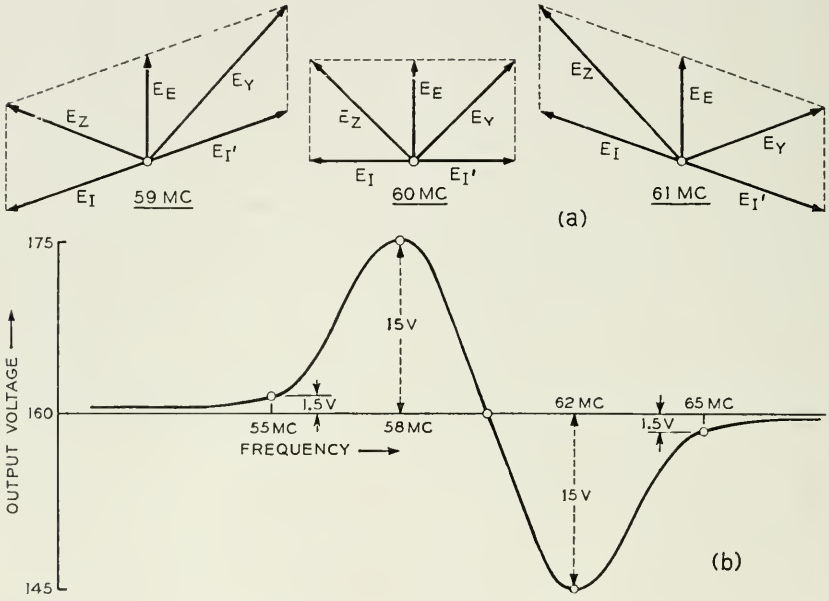


Fig. 69.—Operation of the AFC circuit—vector relationships in input circuit and output voltage vs. frequency characteristic.

to the maximum frequency rate of change which this circuit must control. The d-c amplifier shown is normally biased to operate at maximum gain consistent with stability, to produce the maximum sensitivity to frequency change and to accordingly achieve the least threshold deviation from the ideal tuning condition. Provision for disabling the AFC circuit is included to enable initial manual adjustment of the beat oscillator repeller potential. With the circuit shown here, failure of the AFC circuit proper will result in the return of the repeller potential to that value originally selected by initial tuning and further manual control may be used.

It is often convenient to describe the effectiveness of an AFC circuit in terms of its "pull in" range and its "hold in" characteristics. "Pull in"

will be defined as the ability of the AFC circuit to restore proper tuning conditions with sudden application of the signal. The "hold in" characteristic of the AFC system will be defined as the ability of the circuit to maintain proper tuning conditions as slow changes occur in the frequency of the control signal. In the AN/APS-4 airborne radar equipment previously referred to, which employs an AFC circuit similar to the form here described, the "pull in" range is approximately ± 5 mc from the 60-mc midband value and the "hold in" range includes the entire tuning range of the reflex oscillator employed which is of the order of ± 40 mc. This example will maintain the tuning within 0.5 mc of the desired tuning point over the range of conditions encountered in wartime aircraft applications.

In some applications use has been made of a frequency scanning process whereby the AFC output voltage, in the absence of a suitable controlling signal within the IF band, is caused to vary periodically in a saw-tooth fashion thus causing the local beat oscillator frequency to vary, sweeping across the complete tuning range of the receiver. When the desired signal frequency is produced the AFC then functions in the normal manner. This form of circuit was employed in certain radar equipments developed during the early part of the war and a somewhat similar oscillatory AFC circuit has been employed in connection with later developed thermally tuned reflex oscillators and reported elsewhere.[13]

An automatic frequency control unit designed in connection with the AN/APQ-7 radar bombing equipment which operates at 10,000 mc is illustrated in Fig. 70 and Fig. 71. The basic operation of this equipment example is similar to the d-c amplifier type previously described but includes certain modifications important for this particular application. In this circuit the plate potential of the first IF stage of the AFC unit is obtained as a positive pulse from the transmitting modulator thus enabling the AFC circuit only during the short interval of time encompassing the outgoing radar pulse. This arrangement assures that no detuning of the receiver will result from spurious or nearby signals after the radar pulse has been transmitted. The rectifier elements here consist of triodes operating near plate current cut-off which results in improved linearity of detection. The d-c amplifier portion of this AFC circuit is arranged somewhat differently from the example previously discussed, including in this case balancing controls to account for tube and circuit variations. At the condition of resonance, in this case 60 mc, the voltages applied to each grid of the amplifier are equal and the net repeller potential is determined entirely by the manual control value. The overall output range of voltage for this circuit is ± 20 volts, which in this application represents a ± 40-mc frequency change for the associated

[13] "Considerations in the Design of Centimeter-Wave Radar Receivers", Stewart E. Miller, *Proc. I. R. E.*, Vol. 35, No. 4, April, 1947.

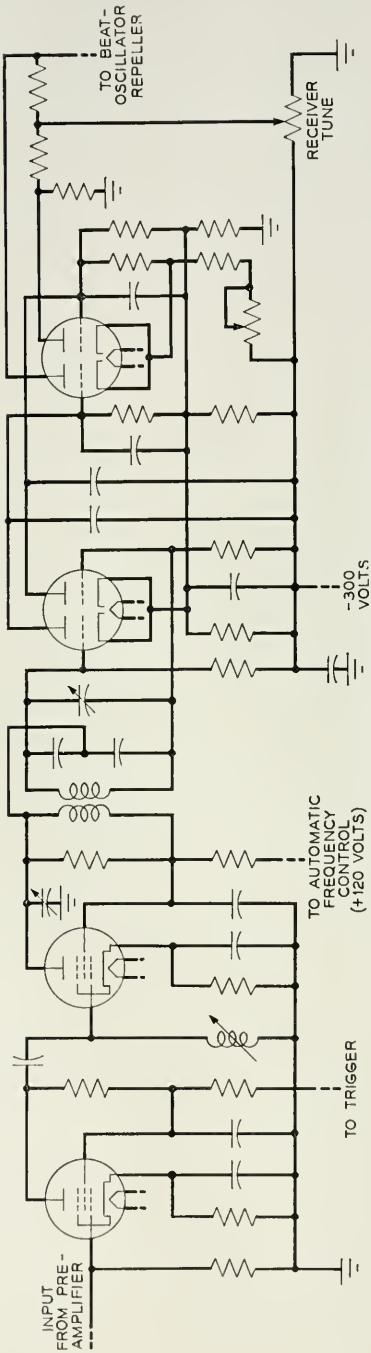


Fig. 70.—AFC circuit schematic for AN/APQ-7 radar system.

reflex oscillator. Variations in the amplitude of the controlling signal are of less importance by virtue of the biasing action of this d-c amplifier circuit. The performance of this design includes maintenance of the tuning point of this receiver to within $\pm$ 0.25 mc of the center of the IF band which in this case is 60 mc.

2.72 *Automatic Gain Control*

Automatic gain control (AGC) of a selected radar signal is often required in military radar systems employed for fire control or aircraft interception purposes. In the case of the common search radar system, AGC is seldom required. The radar receiver AGC function is quite similar to that required of this circuit in the usual radio communication system, i.e., automatic am-

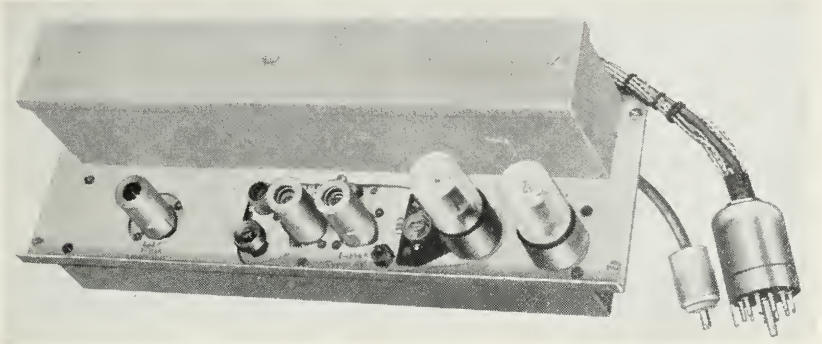


Fig. 71.—AFC component design as employed in AN/APQ-7 airborne radar system.

plitude stabilization of the desired signal. For the radar receiver case, however, the desired signal must be selected on a time interval basis.

In the usual type of automatic tracking radar system, the target is selected by manual alignment of a range and/or bearing "gate". This gating process is essentially a modulation process by which the complete received radar pulse signals are modulated with a rectangular pulse synchronized with the outgoing radar pulse. The modulating pulse or gate has a finite amplitude only over the time interval under observation. In this manner all received information, except that occurring during the selected time interval, is effectively rejected and the automatic gain control circuit operation is entirely defined by the data present during this time interval.

The remainder of the automatic gain control circuit is concerned with the measurement of the amplitude of the selected signal, usually by a peak voltage measurement, the averaging of this measurement over a convenient time interval, and the production of a suitable gain control voltage to be impressed upon the radar receiver IF or video amplifier circuit. The detailed

AGC circuit design is dependent to a major extent upon the dynamic characteristics of the associated automatic tracking device. The subject of automatic tracking design principles cannot be reviewed here and accordingly more detailed AGC design consideration must also await inclusion in such a future report.

2.8 *Radar Receiver Power Supplies*

The remaining components of the radar receiver to be here reviewed consist of the power supplies necessary to produce the various d-c potentials as required for the operation of the electronic components of the receiver. The principal design problems associated with these components arise from the relatively poor stability characteristics of the prime sources of power available at the military scene and the rather severe output voltage requirements to adequately serve the precision nature of radar reception, display, and measurement. The supply voltages necessary for a military radar receiver range from low bias potentials upward to 5000 volts for cathode-ray tubes and TR application with both polarities often required.

2.81 *Primary Power Sources*

The characteristics of the primary source of power available for the military radar system are dependent on the area of use of the equipment. In the case of mobile ground radar installations, the gasoline engine driven generator represents the typical primary power source. In the case of large mobile radar systems it is customary to employ 115-volt—60-cycle primary power, while in certain more portable designs 115-volt—400-cycle primary power has proven satisfactory. The frequency and output voltage of a gasoline engine driven alternator cannot be maintained within the narrow limits desired by the radar receiver and here the major burden of precise voltage regulation must be carried by the electronic regulated power supply within the radar receiver.

For naval vessel radar installations 115–230-volt—60-cycle primary power is commonly available on the larger vessels. For PT and similar smaller craft, certain radar installations have been employed operating from 24–48 volts direct current with motor generator sets supplying 60-cycle or 400-cycle power for the radar system. In the case of undersea craft, the storage battery is employed as a primary source of power and motor generator sets are employed to obtain 115 volts, 60 cycles in most instances.

The primary source of power for aircraft radar purposes is either a low voltage (27 volts) d-c generator driven by the aircraft engine or an alternator similarly driven. If d-c power is available, an additional motor generator set may be employed to furnish the 115-volt—400 to 800-cycle power for the radar equipment use. Voltage regulators of the carbon pile compres-

sion or electronic types are usually employed here, resulting in a nominal ± 3% voltage stabilization. The extreme variable electrical loads imposed upon the aircraft power system by electrically operated gun turrets and other combat equipment result in increased emphasis being placed on adequate regulation capabilities of the radar receiver power supplies. In addition, the ever-present requirement of minimum weight for aircraft equipment results in motor generator designs employing a minimum of magnetic material and usually results in a variation in output voltage wave shape with load. This factor must also be considered in the detailed design of the aircraft radar receiver power supplies.

During the initial airborne radar development program, the power frequencies in common use were 400 cps and 800 cps. The British use of direct coupled aircraft engine driven alternators produced variable frequency output voltages ranging from 1200–2400 cps dependent on the engine speed. In connection with the electronic warfare standardization program for equipment to be used jointly by our allies, the aircraft radar system was required to operate over the entire range of power frequencies extending from 400 cps to 2400 cps. All of the airborne radar equipment developed during the later half of World War II was designed to meet these variable power frequency requirements.

2.82 *Low Voltage Power Supplies*

The electronic regulated power supply has been universally employed to furnish the stable low voltages as required by the radar receiver. Here the output voltages required extend from 50 volts to 600 volts with maximum direct current required extending upward to 500 ma.

The basic electrical circuit arrangement of such a power supply is shown in Fig. 72. In this circuit, the regulating element consists of a variable series impedance, furnished in the form of a vacuum tube and resistance combination, whose magnitude may be controlled electrically from an error signal associated with the output voltage of the power supply and a reference voltage. The control circuit consists of a bridge network which includes a constant voltage gas-discharge tube as one element. A d-c amplifier is connected across the output terminals of this bridge circuit and serves to amplify the error signal for use in the regulating element. If the output voltage of the power supply varies from the desired value for any cause, the error signal appears at the output terminals of the bridge circuit, due to the effective unbalance of this circuit at all voltage levels except the reference value. The error signal after sufficient amplification is impressed upon the grid of the series regulating tube with a polarity such that a corrective impedance variation results. The degree of regulation obtainable is a function of the loop gain provided and the absolute stability of the output voltage

is determined primarily by the constancy of the reference voltage derived by the use of the gas-discharge tube. By incorporating wide frequency band characteristics to the loop gain elements, the maximum rate of change of regulation can be extended, and this circuit becomes very effective in reducing the fundamental and harmonics of the primary supply voltage. The effective impedance of a radar receiver power supply of this type is of the order of one ohm, a factor of extreme importance in reducing the unwanted interaction between the various receiver components by coupling due to this common impedance. Other variations in circuit arrangement for regulated power supplies are occasionally employed, the most common of which involves the use of a vacuum tube as a shunt regulating element as contrasted with the series arrangement shown here. In certain low-current

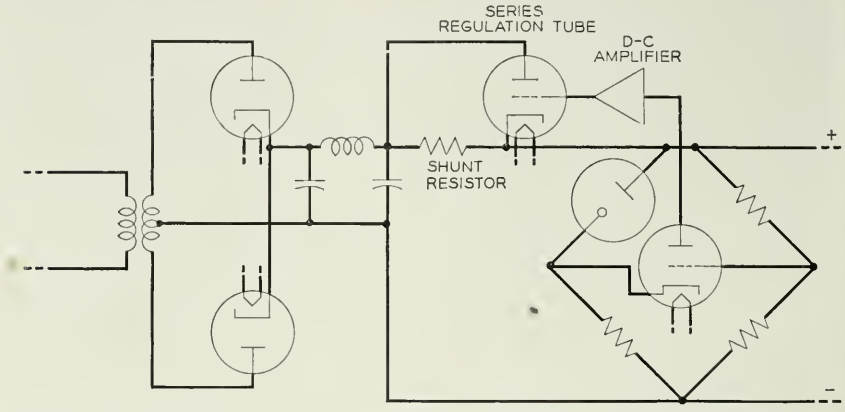


Fig. 72.—Simplified circuit schematic of low-voltage power supply—Series regulation type.

applications the use of gas-discharge tubes of constant voltage characteristics are occasionally employed in a shunt circuit configuration.

Figure 73 illustrates the equipment arrangement of a radar receiver power supply as employed for an airborne radar bombing system. In this example is included one non-regulated and three regulated rectifier power supplies with output voltages of +600, +300, +120 and −300 volts available for the radar receiver. The forced ventilation feature shown here is required to prevent extreme temperature rise of the components under high altitude conditions in the presence of considerable heat dissipation by the rectifier and series regulating circuits. Each power supply is designed as a separable subchassis within the over-all enclosure to provide for ease in manufacture and testing of the unit. The weight of this complete unit as installed in a military aircraft is approximately 50 pounds.

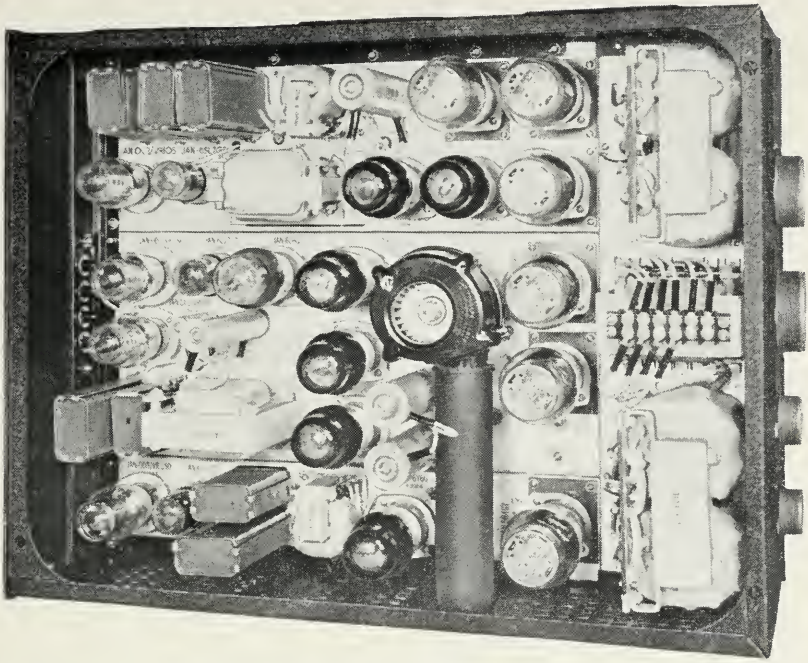


Fig. 73.—Low-voltage power supply for AN/APQ-7 airborne radar system—Mechanical features.

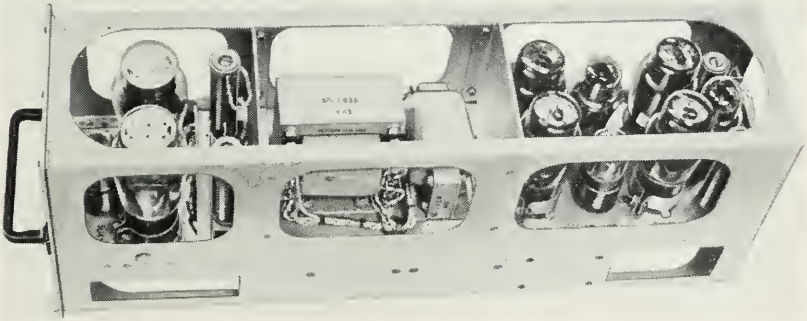


Fig. 74.—Low-voltage power supply for AN/APQ-5 low altitude radar bombing equipment.

Figure 74 indicates the mechanical design features of a typical airborne radar receiver power supply of the series regulated type as employed in the AN/APQ-5 radar bombing equipment. This example illustrates the emphasis which is placed on compactness and lightweight construction of airborne radar components.

To obtain the maximum dependable performance from the various power transformers and coils employed in the radar receiver power supplies under the severe military field conditions, a considerable development program was carried on throughout the past war. At the beginning of this program the only available method of insuring adequate transformer winding insulation under extreme humidity conditions involved the sealing of the structure in a metal container. For aircraft service, where weight is of prime concern, this added weight could not be tolerated so that here an open type of structure was extensively employed, with the protection to the winding being furnished in the form of several coats of varnish followed by an enamel overcoat. With the increased emphasis on high-altitude operation of military aircraft and the rapid temperature and pressure changes involved, a development program was instituted to improve the open-type transformer sealing process. The result of this program has produced the Flexseal process whereby the service life of this type of power transformer has been increased as much as ten times that obtained with the varnish process formerly employed. This process involves a multiple-dip varnish coating method where the varnish is thickened by the addition of talc, a very fine low-gravity wettable inert filler. This process results in the formation of a relatively thick plastic shell which completely surrounds the transformer structure. One of the features of this process is found in its simplicity, whereby no special equipment was required to carry out the procedure, an important factor during a wartime production program. Figure 75 illustrates a number of typical open-type power transformers which were employed on various military radar projects, all of which incorporate the Flexseal treatment for improved service life.

2.83 *High Voltage Power Supplies*

The high voltage, as required for radar receiver cathode-ray tube indicator purposes, varies from 2000 volts to 5000 volts, and for the TR tube keep-alive potentials of the order of 1000 volts must be provided. In these cases, however, the d-c current requirements are quite small and generally no regulation means are required for stabilization of the voltage, the stability of the primary source of power usually being sufficient.

The design problems encountered in this type of power supply are concerned primarily with the requirements of reliability of operation under severe military operating conditions, and further require that only circuit elements having well-defined factors of safety be employed in such applications.

Figure 76 illustrates a number of typical high-voltage transformer designs which have been employed in military radar systems during the past war period. Both air insulated and oil immersed types of structures are shown.

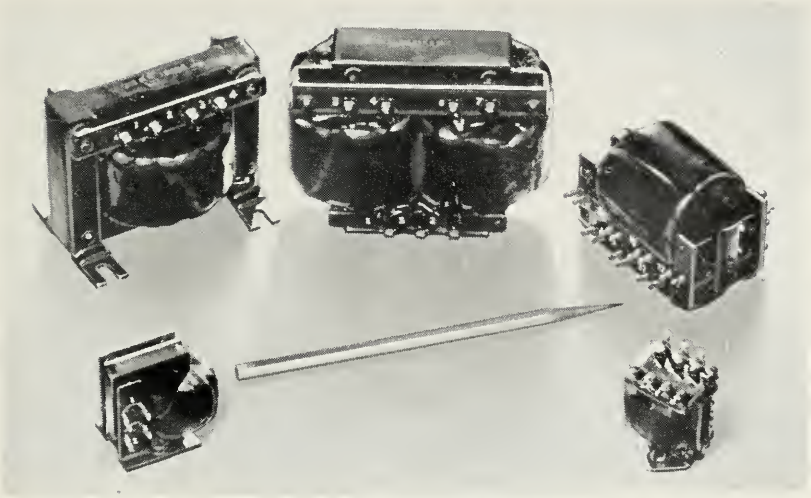


Fig. 75.—Flexseal treated open-core type power transformers as developed for airborne radar application.

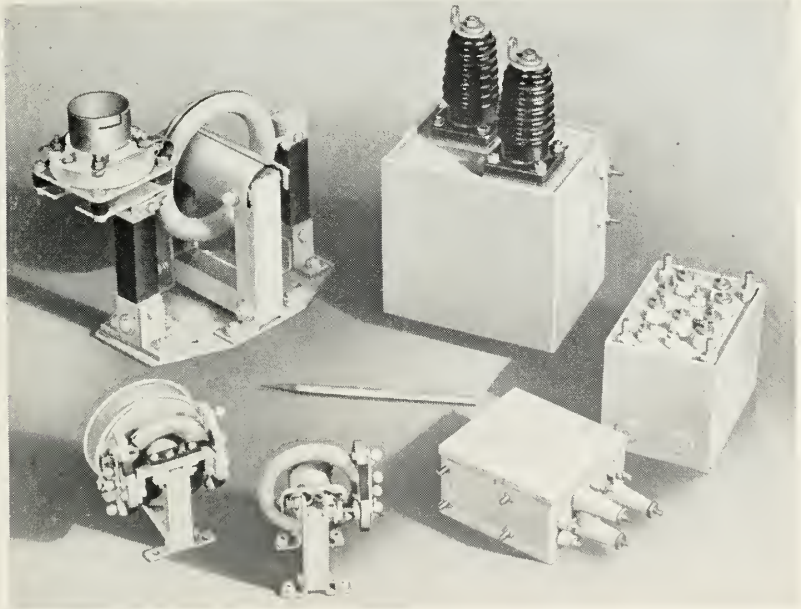


Fig. 76.—Power transformer designs for high voltage radar power supply application—Air-insulated and oil-filled types are included.

The use of air insulation in a high-voltage transformer results in a relatively low coupling coefficient between primary and secondary windings and

accordingly restricts the range of power frequencies over which satisfactory regulation and operation can be expected. With the emphasis on power supply frequencies extending from 400 cps to 2400 cps for aircraft purposes, this air insulated type of structure was abandoned in favor of oil immersed types. The primary disadvantage of the oil immersed type of transformer is the increased weight of the unit.

A high-voltage power supply design as produced for airborne radar system application is shown in Fig. 77. In this example +4900 volts is supplied for cathode-ray tube indicator purposes and −1000 volts is available for the TR tube keep-alive circuit. This unit employs an air insulated type of

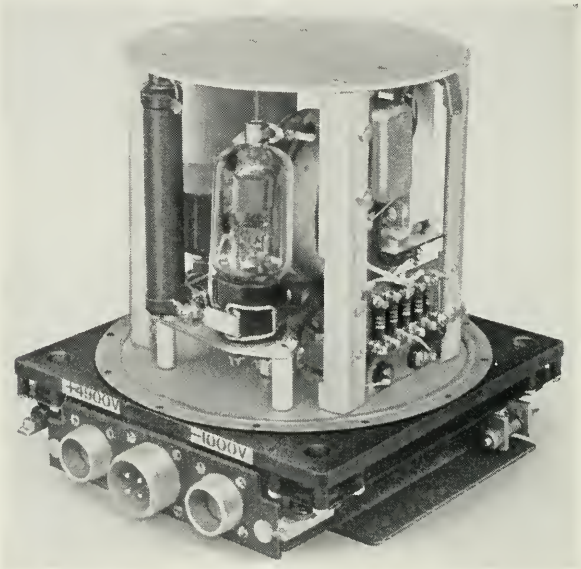


Fig. 77.—High-voltage power supply for airborne radar receiver application—pressurized type.

high-voltage transformer and by the use of a hermetically sealed enclosure operated at sea-level atmospheric pressure, satisfactory performance at high altitudes is realized.

Conclusion

The complete technical story of radar is of a magnitude comparable to a detailed report of the military campaigns of this past global war. During this period, the Bell Telephone Laboratories developed for manufacture more than 70 specialized radar systems for the Armed Services. It has been possible here only to review the major technical considerations which in-

fluence design of a military radar receiver and to present a few radar circuit and equipment illustrations of the specialized technology that has resulted. It is a tribute to the ingenuity and industry of the workers in the radar field that this technology, developed under extremely accelerated and difficult conditions, will have a permanent value in future communication systems design.

The material used in this paper represents the concrete contributions of countless individual workers within and without the Bell Telephone Laboratories. The equipment products illustrated are the product of concentrated effort on the part of the staff of the Western Electric Company who produced these specialized and complex radar systems in quantities and within schedules necessary for the successful prosecution of the war. It is regrettable that it is an impossible task to assign individual credit for these specific military radar contributions.

High-Vacuum Oxide-Cathode Pulse Modulator Tubes

By C. E. FAY

Introduction

IN PRACTICALLY all pulsed oscillators such as those used in radar, some means must be provided to apply the pulse voltage to the oscillator circuit. In many early radars, a high-vacuum modulator was used for this purpose. The pulse was generated at low power level and then amplified by means of one or more stages employing high vacuum tubes. The final stage was required to block, or cut off the d-c supply voltage with no pulse applied, and to permit as much as possible of the d-c voltage to appear on the oscillator during the pulse. Since most radar oscillators operate at pulse voltages of from 5 to 20 kv and require currents of several amperes during the pulse, the requirements of the modulator tubes are quite severe. Standard transmitting[1] tubes were used at first, the higher power tubes having the necessary voltage rating and having in general a fair amount of cathode emission. Tubes were operated in parallel to provide the required amount of current. Practically all of these tubes were of the thoriated tungsten filament type. For example an early army radar, the SCR268,[2] employed 8 tubes in parallel having a total filament power of 1040 watts to provide a pulse current of about 10 amperes. The use of such equipment in portable or airborne service would be obviously impractical because of the large power consumption, bulk, and weight. In an attempt to provide tubes more suited to this type of service, those described in this paper were developed.

Tube Requirements

The function of the high-vacuum modulator tube essentially is to act as a switch to turn the pulse on and off at the transmitter in response to a control signal. The best device for this purpose will be the one which requires the least signal power for control and which allows the transfer of power with the least loss, from the transmitter power source to the oscillator.

If the oscillator must be supplied with a pulse of voltage E_p and current I_p,* or power E_pI_p, then the voltage which must be supplied by the transmitter power supply will be $E_b = E_p + e_p$, Fig. 1, if e_p represents the voltage

* It is assumed here that the pulse is rectangular in shape. This is usually the desired shape and is fairly well approximated in most cases.

drop in the modulator tubes necessary to allow current I_p to pass. The plate efficiency of the modulator is then simply $\frac{E_p}{E_b}$ and the power dissipated in the modulator tube plate is $I_p e_p$ during a pulse. The average power dissipated in the plate is then $I_p e_p$ multiplied by pulse length and by pulse frequency. The heat storage capability of the plate is ordinarily great enough that the average power is all that needs consideration.

The conditions imposed on the modulator tube are somewhat analogous to those of a class C amplifier at low frequency. The main difference is that the angle of operation is very small, and there is usually no appreciable backswing of plate voltage since the load is essentially a resistance. Typical

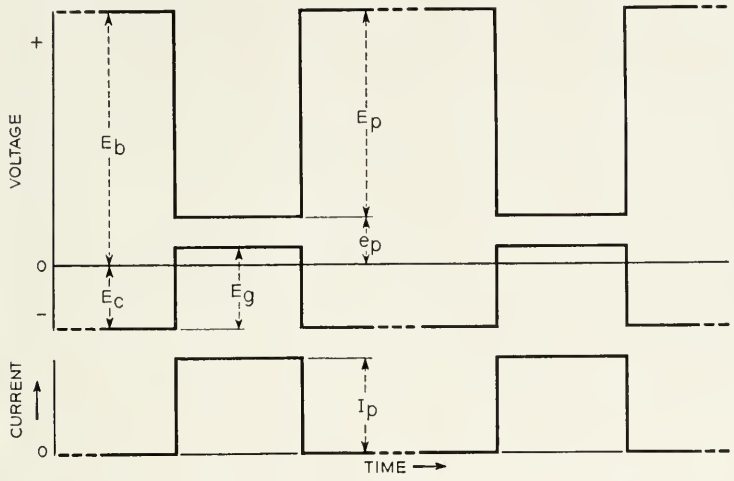


Fig. 1—Current and voltage relations in a pulse modulator tube.

modulator circuits are shown in Fig. 2. It is sometimes found desirable to employ a shunt inductance across the oscillator in the interest of a sharp cutoff of the pulse on the oscillator, particularly where capacitances to ground of various circuit components are appreciable. This results in an additional current demand on the modulator tube since the current through the inductance must also be supplied. The oscillator is often coupled to the modulator circuit by means of a transformer in order that desirable impedances are realized in each circuit.

Design Considerations

It was apparent on first consideration of the high-vacuum modulator problem that use of oxide coated cathodes would be of enormous advantage

in keeping power requirements down. Heretofore the use of oxide cathodes in high voltage power tubes had been found very difficult, particularly where filamentary cathodes were employed. Any spark or momentary discharge in operation usually resulted in the burning out of the relatively fragile filaments. This result was caused mainly by the fact that a considerable amount of energy was of necessity available from the power supply equipment. However, in pulse service it is possible to limit the amount of energy available so that a momentary tube breakdown will not result in damage to a reasonably rugged equipotential cathode. Also, in the interest

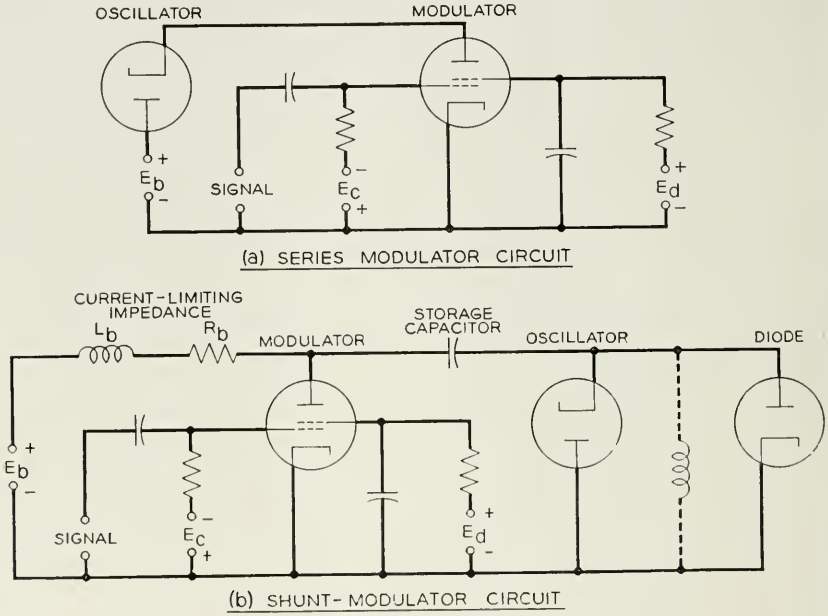


Fig. 2—Typical pulse modulator circuits.

of conserving control power it is desirable to build high perveance tubes which require very close control-grid to cathode spacings. This is much more easily accomplished with rigid cathode structures rather than filamentary cathodes, especially for service conditions under which extreme shock and vibration may be encountered.

Conservation of drive power requires that the modulator tube have high power-gain. This is most easily provided in the tetrode which provides a high over-all amplification factor with reasonable drive characteristics. Lining up the control-grid and screen-grid wires is of course advantageous in the interest of minimizing screen dissipation and getting the largest possible portion of the cathode current to the plate. The control-grid to

plate capacitance is of little importance as long as it does not store much energy, there being little chance for oscillation in such a circuit. While it is desirable to operate with as low a minimum plate voltage as possible, it is of little additional advantage to bring the plate voltage below the screen voltage if the screen voltage is about 1000 volts and the supply voltage 15 kv. It was therefore thought permissible to increase plate to screen spacing beyond the optimum for best characteristics in the interest of high voltage and screen dissipation ratings.

The insulation in the tube between plate and other electrodes must be capable of withstanding the full supply voltage plus a comfortable margin. This dictates that if internal insulators are used they must have long path and that the bulb must have sufficient length to prevent flash-over externally.

The 701A Vacuum Tube

At the time of this development a tube was very urgently needed for a Navy radar application.[3] Since speed was of prime importance it was decided to take parts of a standard oxide-cathode beam-power tetrode, Western Electric 350A, and mount them in a structure capable of withstanding the required voltage; 12 kv in this case. Accordingly a cruciform structure was designed in which four sets of 350A electrodes were mounted on ceramic members attached to a molded glass dish-stem as shown in Figure 3. The four cathodes have a total coated area of approximately 14 square centimeters. A molybdenum plate of cruciform section mounted from a lead-in at the top of the bulb was used. This construction eliminated internal insulators between plate and grids other than the bulb. The control-grid of the 350A is normally gold plated to inhibit primary emission. This feature was retained in the 701A and the screen-grid also gold plated. The plate to screen-grid spacing was increased over that normally used in the 350A in order to allow somewhat better cooling of the grids and to allow greater clearance for high voltage reasons. This made the characteristics depart from good "beam tube" performance but at the high voltage condition of operation this was of little consequence. Characteristic curves of the 701A indicating performance under both high voltage and low voltage conditions are shown in Fig. 4. Since no experience was available at the time of this development to indicate what currents could safely be drawn from the cathodes under pulse conditions, the matter of rating these tubes was mainly guesswork since time was not available to await the outcome of life tests under various conditions. The ratings put on the 701A are as shown in Table I.

For the immediate application in hand, which required 12 ampere pulses at about 10 kv, it was decided to specify two 701A tubes operating in paral-

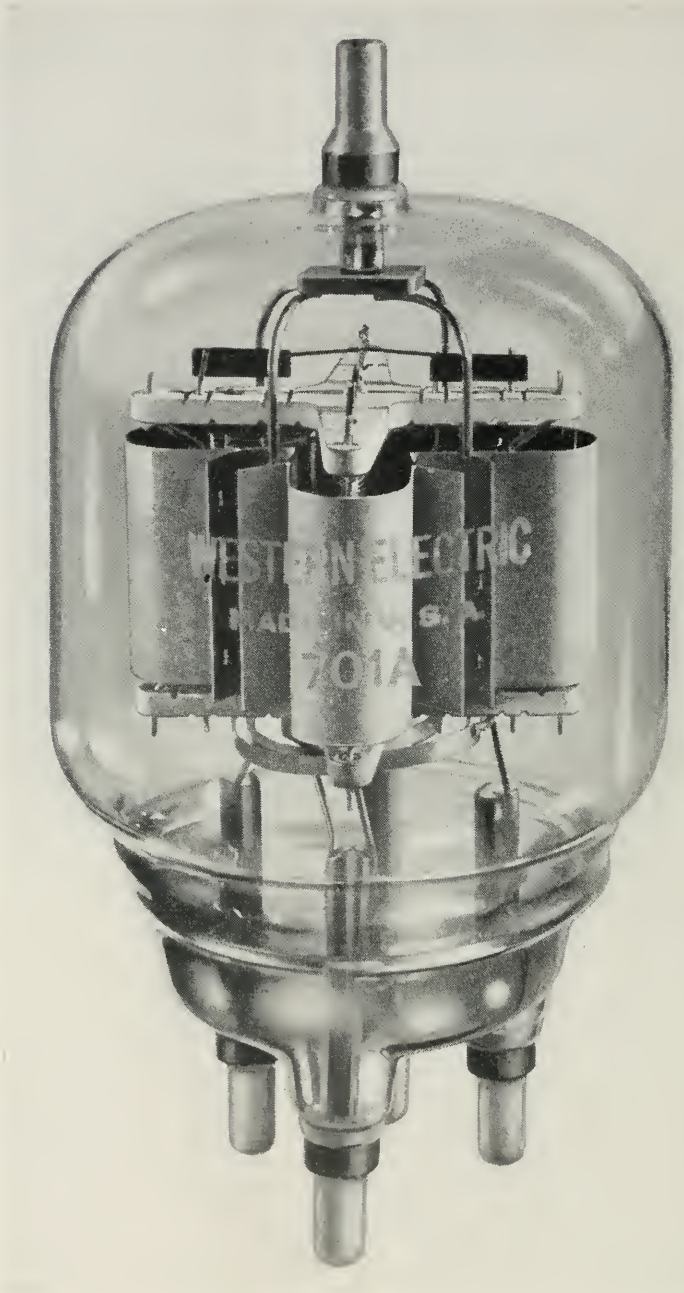


Fig. 3—The 701A vacuum tube.

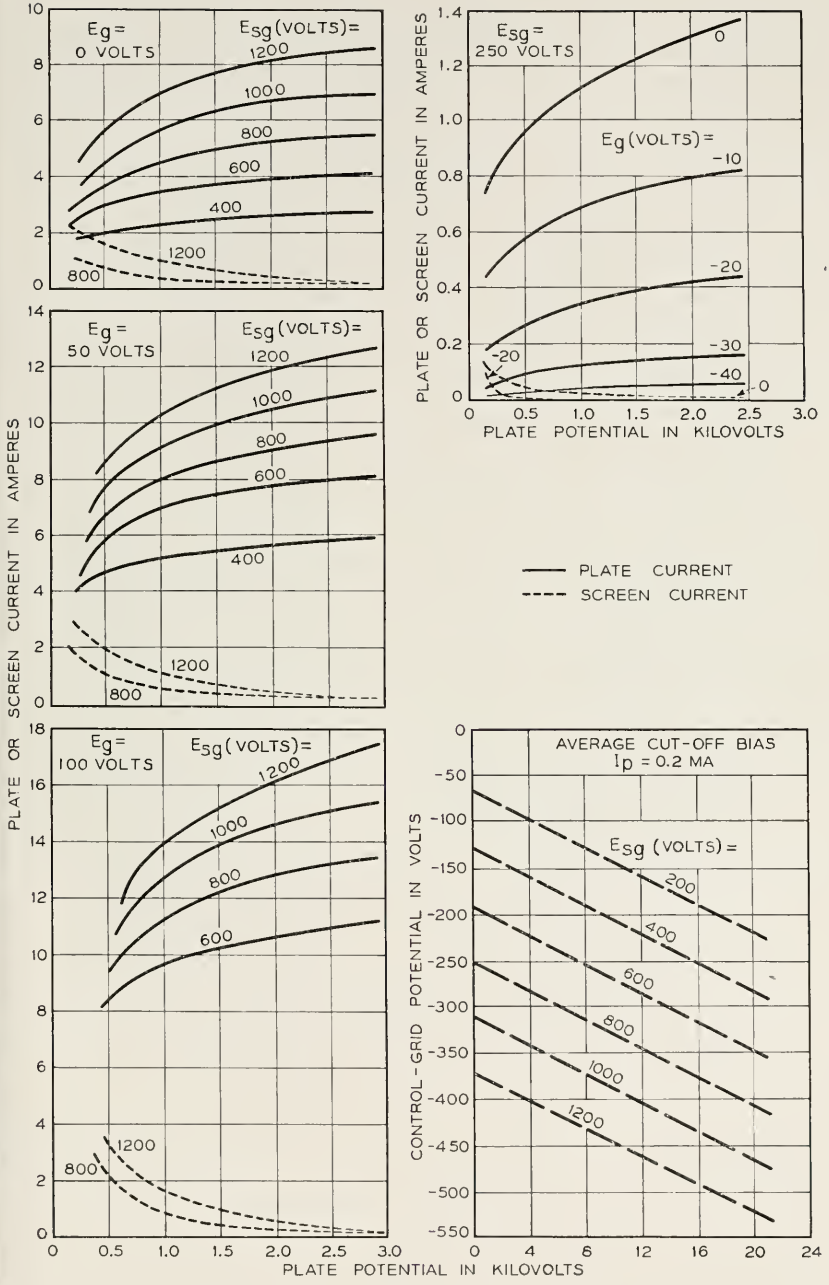


Fig. 4—Characteristics of the 701A vacuum tube.

lel. The pulse in this case was trapezoidal having a base width of about 4μ seconds and a top width of about 1.75 μ seconds; repetition rate was 1600 per second. When the tubes were operated under these conditions there was customarily some sparking within the tube in the first few minutes and then it apparently "aged in," and operated satisfactorily. At the rated heater voltage, the cathodes operated at about 800°C (brightness). This temperature would normally provide a cathode life of more than 1000 hours. Life tests in the laboratory indicated satisfactory performance for about 2000 hours. Reports from the Navy were difficult to obtain but those which were obtained indicated similar results. End of life was caused by both loss of cathode emission and by primary grid emission. Mechanically the tube proved to be reasonably rugged for normal service. However, shocks sustained in shipment of tubes caused mechanical misalignment in

Table I

Table of Ratings of Oxide-Cathode Pulse Modulator Tubes

Tube	Heater Voltage	Heater Current	Peak Plate Voltage	Peak Screen Voltage	Peak Plate Current	Plate Dissipation	Screen Dissipation	Max. Duty Cycle for Peak Plate Current	Capacitances		
									C_{in}	C_{out}	C_{gp}
	Volts	*Amperes*	*KV*	*KV*	*Amperes*	*Watts*	*Watts*		*mmf*		
701A	8	7.5	12.5	1.2	10	100	15	0.005	56	11.5	3.2
715A	27	2.15	14	1.2	10	60	8	0.002	35	7	1.2
715B	26	2.10	15	1.25	15	60	8	0.001	35	7	1.2
5D21	26	2.10	20	1.25	15	60	8	0.001	35	7	1.2
426XQ	8	7.5	25	1.5	20	150	15	0.001	46	7.5	0.6

some cases, indicating a need for a more rugged structure for use in the armed services.

The 715A Tube

The advent of airborne radar made the development of high power lightweight transmitters an urgent requirement. In this case long life was somewhat subordinate to lightweight and small dimensions. Ruggedness was also a requirement. The electronic properties of the 701A tube were well suited for airborne radar but the large bulk was an extreme disadvantage. Work was begun on a tube using the same cathodes as the 701A but having a simpler and more rugged mechanical structure. Out of this evolved the 715A tube. In this tube the cathodes were placed side by side and enveloped by a single control-grid, screen-grid and plate. In order to provide the necessary ruggedness and to keep the grids cool, heavier grid wires were used and they were wound on very heavy supports of high heat-

conductivity material. Both grids were gold plated as in the 701A. All electrodes were mounted between two specially shaped ceramic insulators which provided a relatively long path between plate and grids. This structure is shown in Fig. 5. Heat radiating fins were attached to the ends

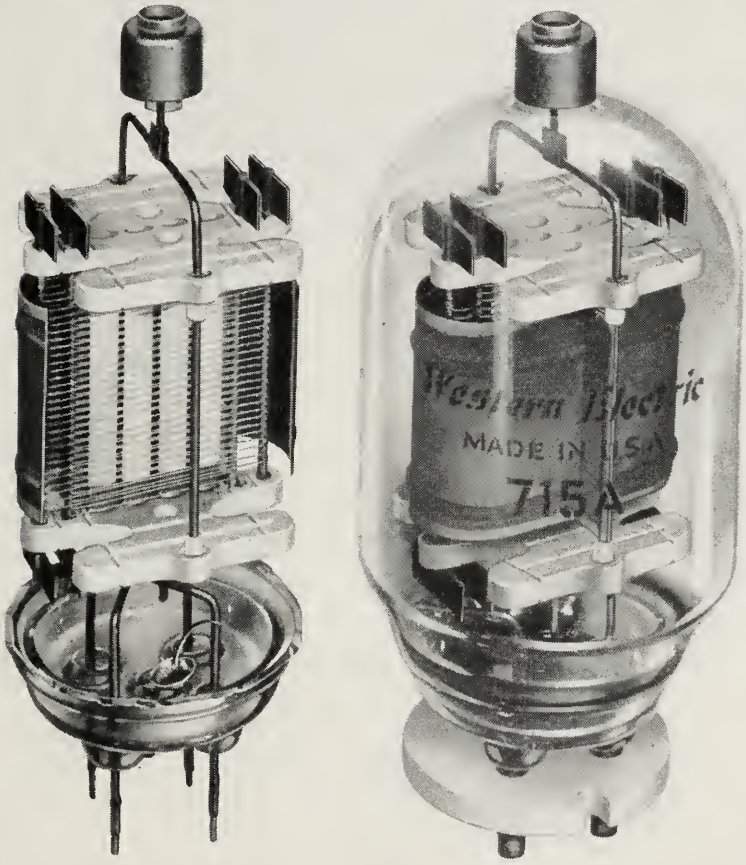


Fig. 5—The 715A vacuum tube.

of the control-grid and screen-grid supports. The plate is molybdenum with zirconium coating on its outside surface. This coating was employed to increase the thermal emissivity of the plate in the interest of a low operating temperature. It also serves to absorb some gas. The cathode, heater and grid terminals of the tube were brought out in the moulded-glass 4-Pin

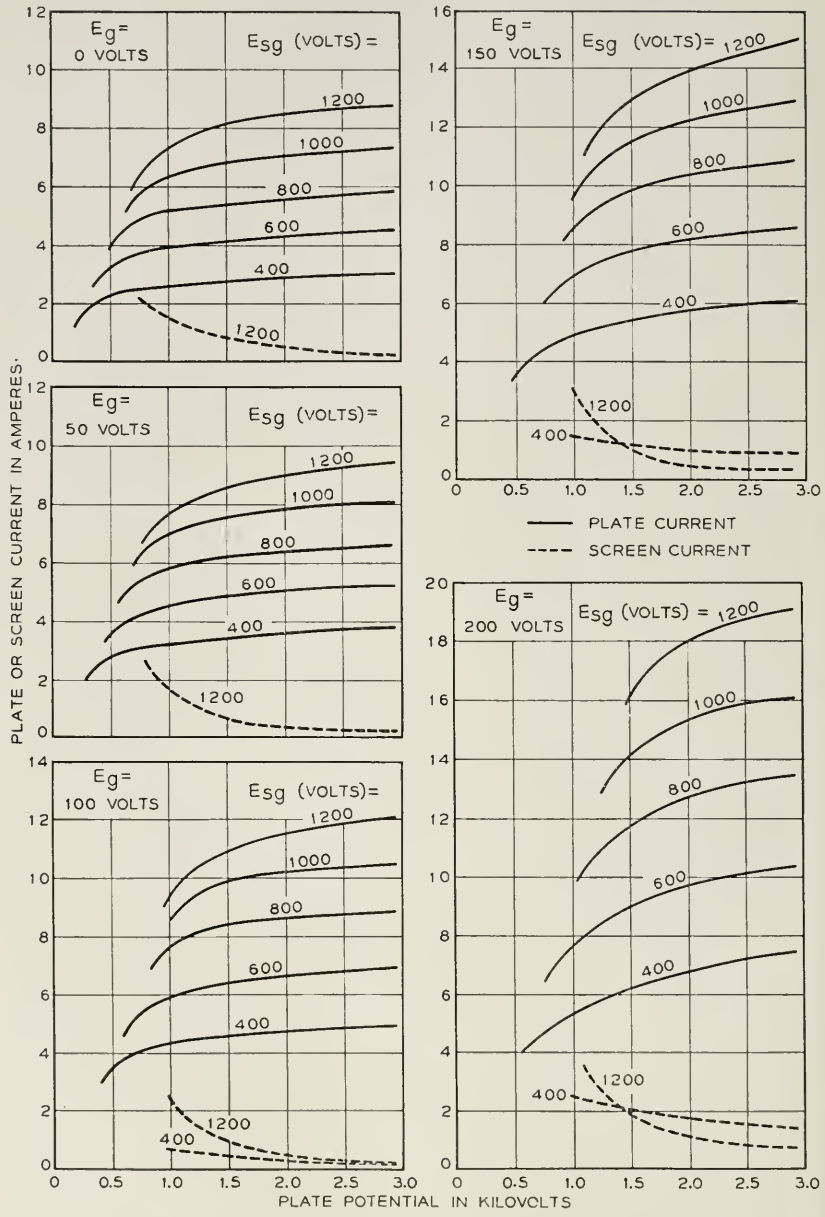


Fig. 6—Characteristics of the 715A vacuum tube.

base, and the plate terminal out the top of the bulb. This provided a very rigid structure which could stand extreme shock and vibration condi-

tions. Although this structure sacrificed something in electronic performance over the 701A it still was quite satisfactory as a pulse modulator.

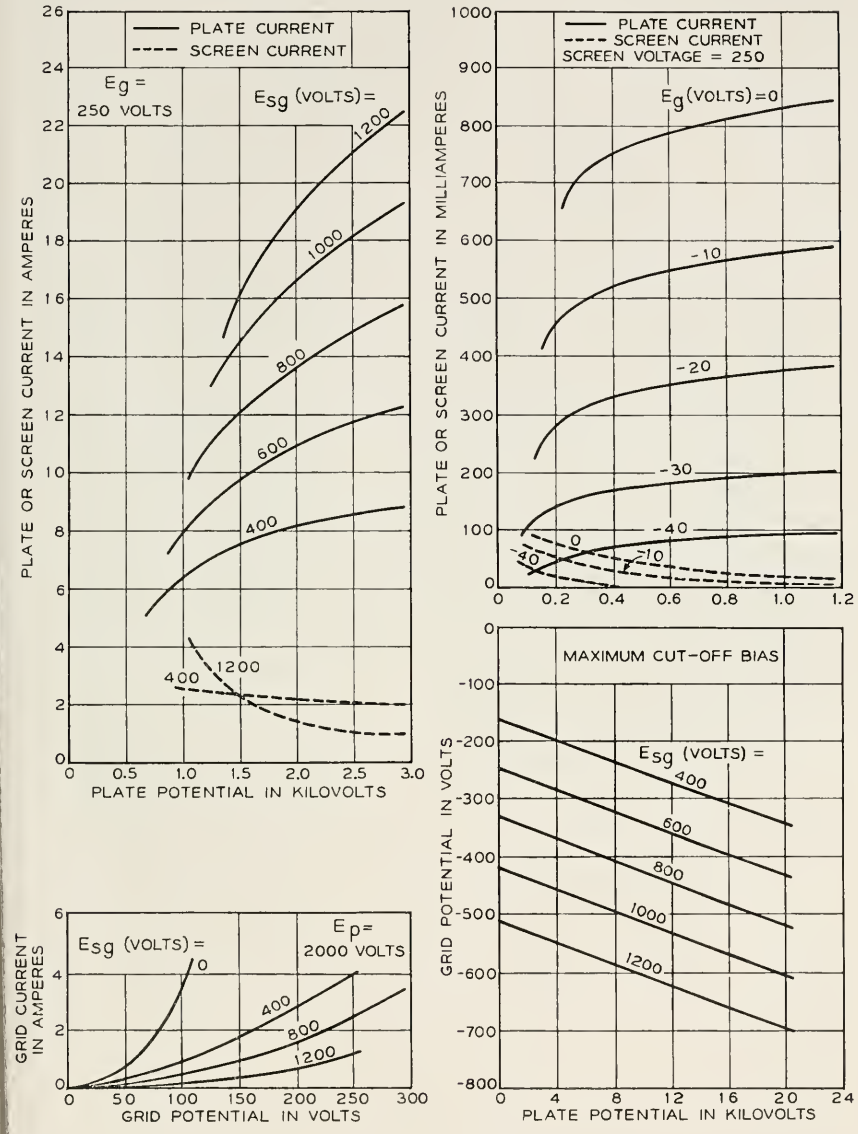


FIG. 6 (Continued)

The characteristics of the 715A tube applicable to both high voltage and low voltage operation are shown in Fig. 6. It was found that in spite of

the internal ceramic insulators, satisfactory operation could be obtained at voltages as high as 15 kv. Since the 715A was designed primarily for airborne applications it was found desirable to design the heater to operate directly from the aircraft's storage battery which was a 24 volt battery.

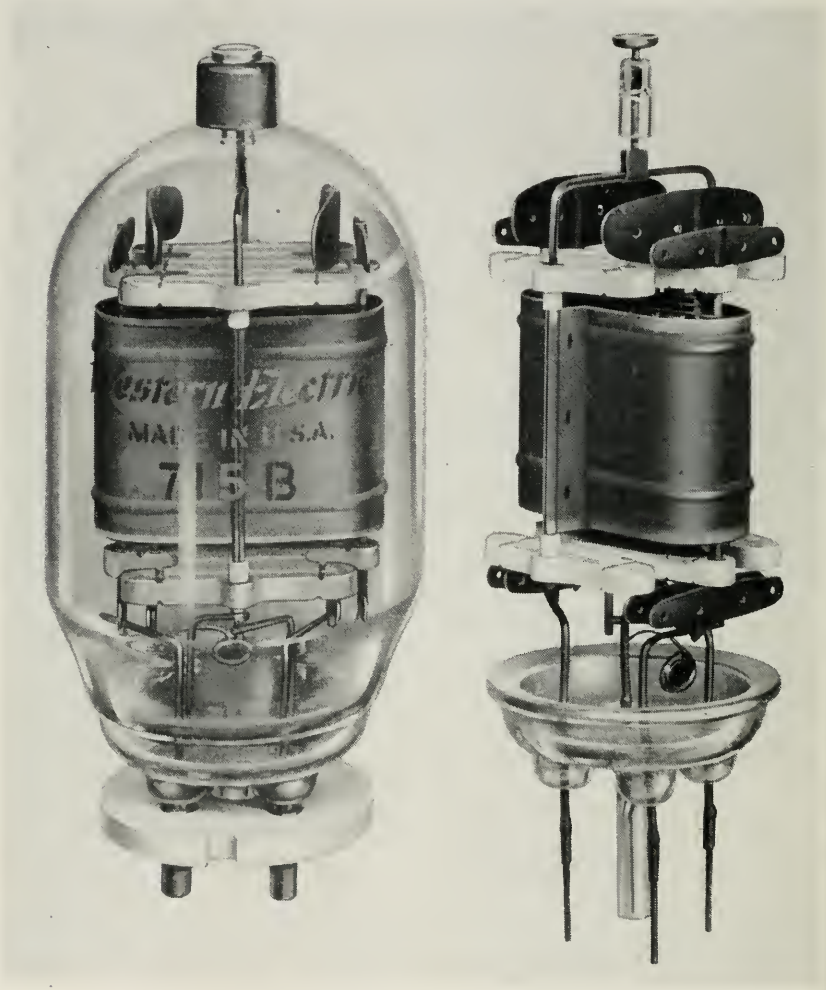


Fig. 7—The 715B vacuum tube.

It was also desirable that the equipment be operable when the charging generator was not running, at which point the voltage might be as low as 22 volts, and also when the generator was charging and the voltage as high as 28.5 volts. This required a compromise in the design of the heater which

resulted in operation of the cathode at somewhat higher than normal temperature under rated conditions. The ratings of the 715A tube are given in Table I.

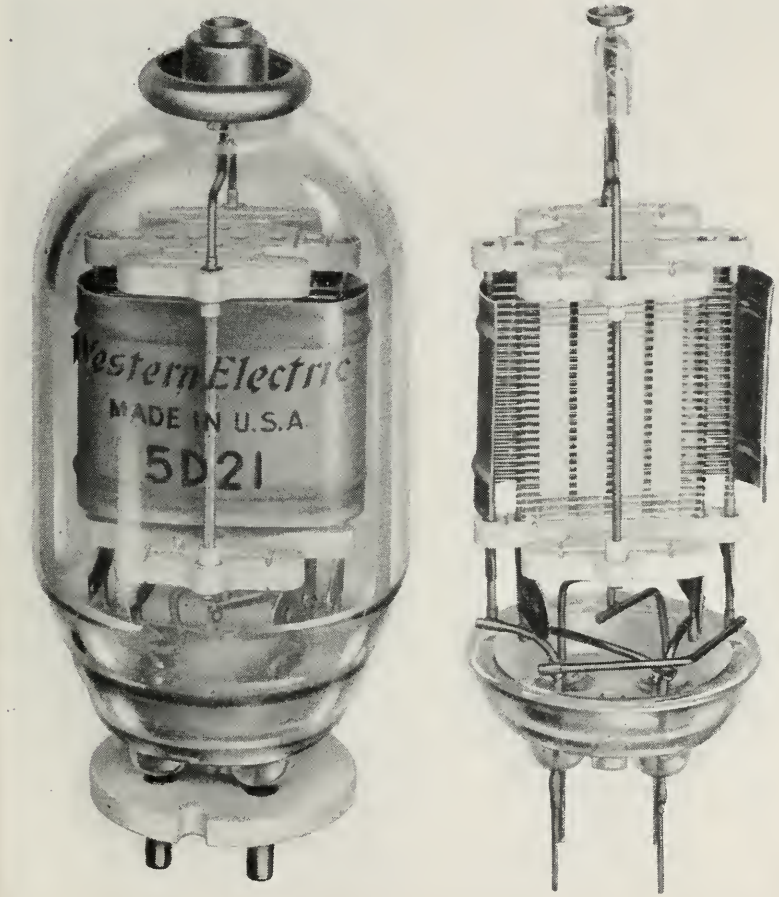


Fig. 8—The 5D21 vacuum tube.

The 715B Tube

Some applications developed which required a peak pulse current slightly greater and of longer duration than that for which the 715A was rated. Meanwhile more experience with the 715A and improvements in processing techniques indicated that a higher peak current rating was justifiable providing the grid temperatures were not increased.

The 715B tube is essentially the same structure as the 715A except that larger radiating fins are attached to the ends of the grid support wires. Figure 7 shows the structure of the 715B tube. The characteristics were identical but the ratings were changed, as indicated in Table I. The life obtained in laboratory life tests under rated conditions averaged between 500 and 1000 hours. Failure was usually caused by grid emission or loss of cathode emission.

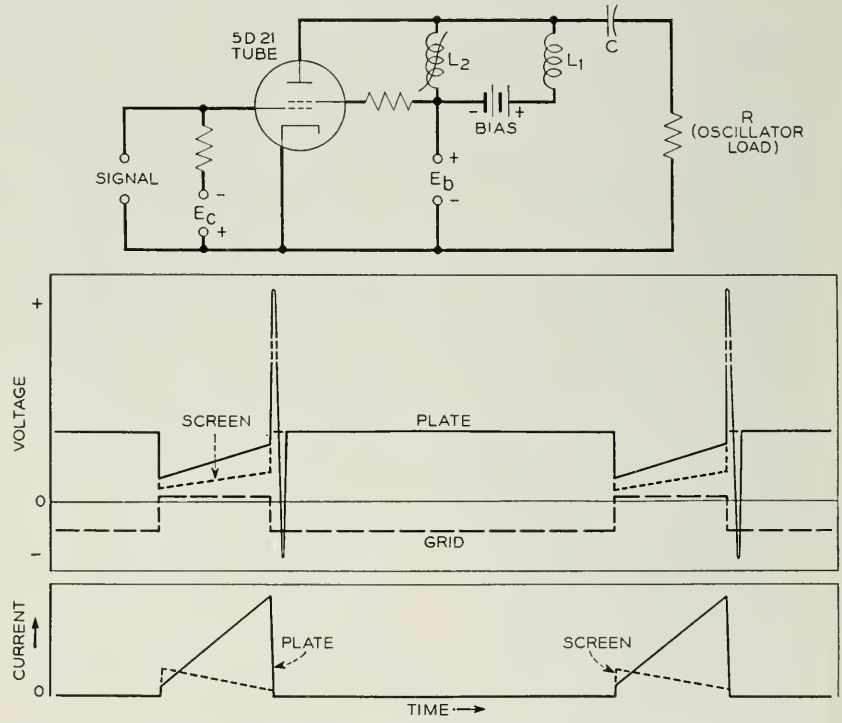


Fig. 9—Non-linear coil modulator circuit with illustration of the current and voltage relations in the 5D21 vacuum tube.

The 5D21 Tube

In response to the demand for further improvement of this structure in its ability to withstand higher voltage, the 5D21 tube was developed, Fig. 8. It is of the same family as the 715A and 715B. It was found that higher voltages could be used if the grid cooling radiators were removed from the top end of the tube. The cooling of the grid was maintained by providing copper wire connections from the bottom ends of the grid support wires to the base seals. This and the use of a specially designed plate terminal cap enabled the voltage rating to be raised to 20 KV.

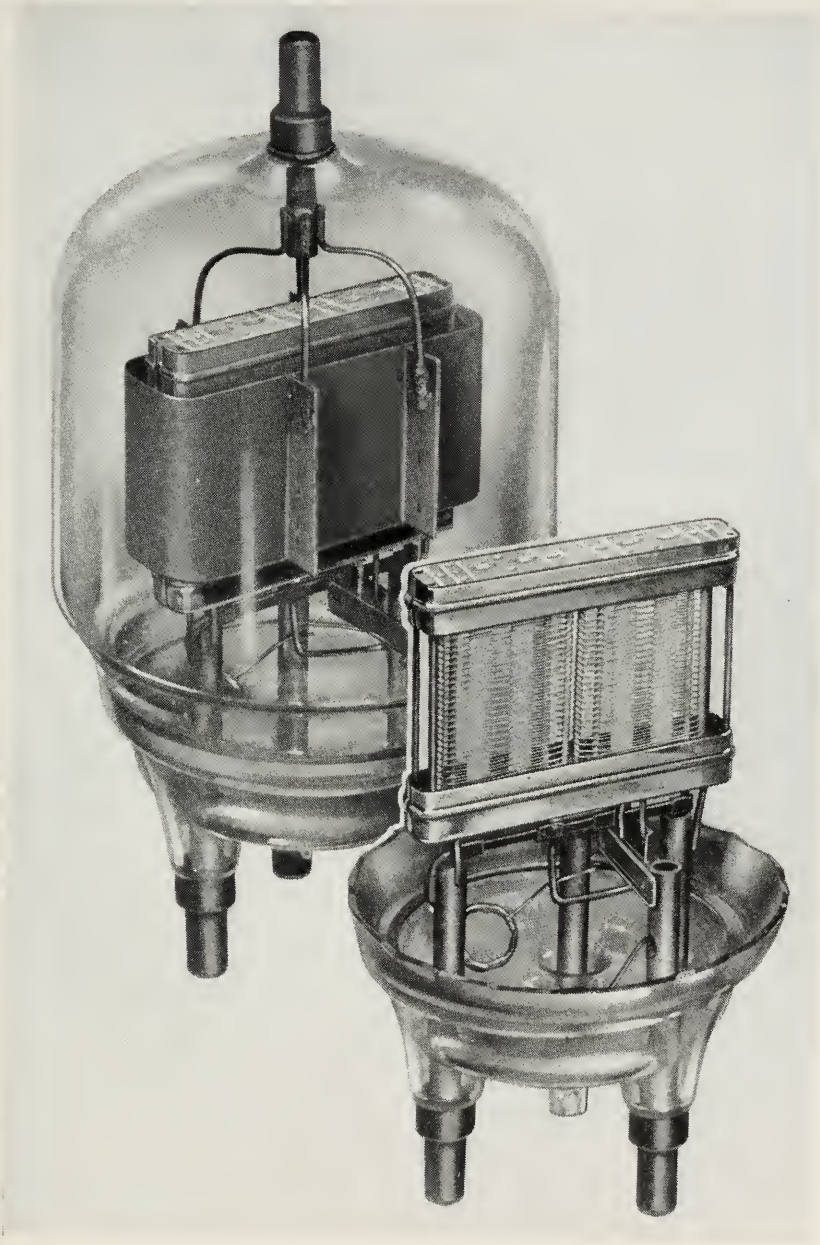


Fig. 10—The experimental 426XQ vacuum tube.

The 5D21 tube also found application as the control tube in non-linear coil type modulators.[4] Here the function of the tube was to permit passage

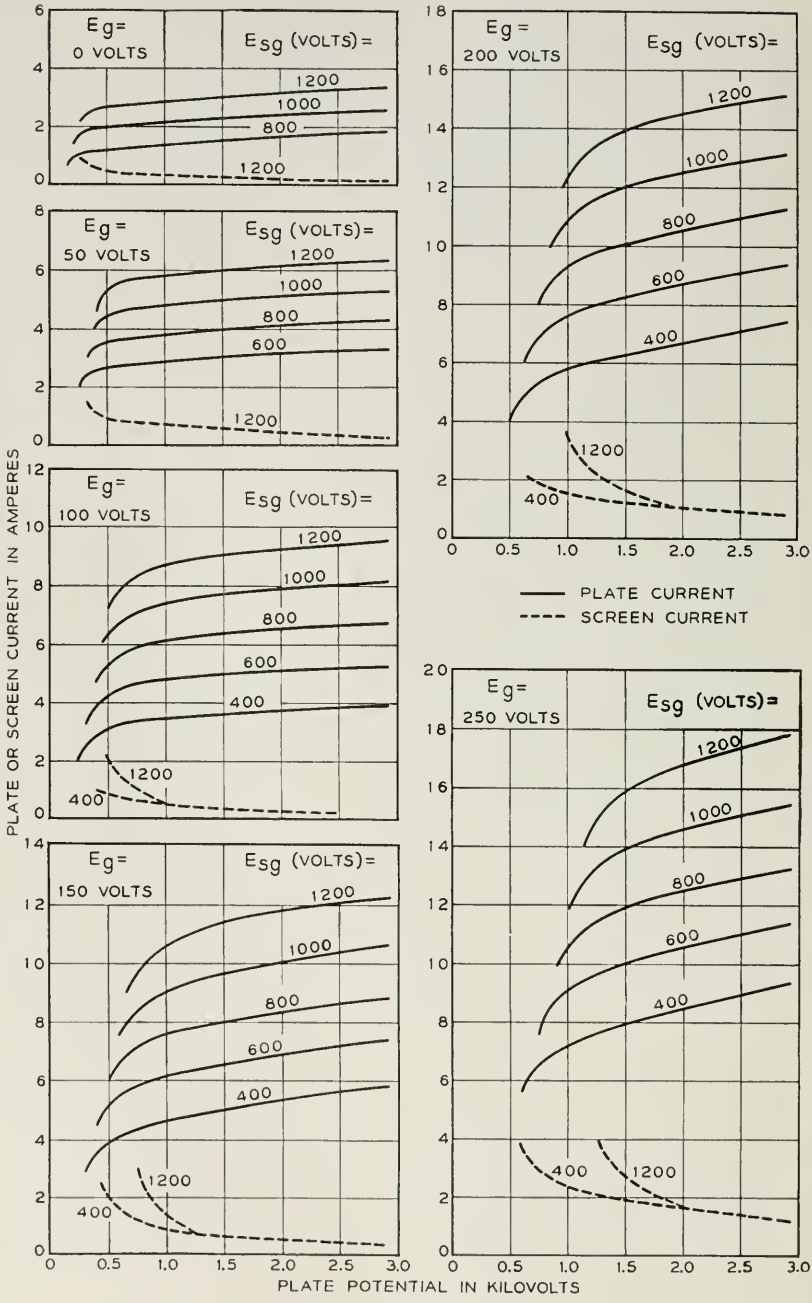


Fig. 11—Characteristics of the 426XQ vacuum tube.

of a moderately high current through an inductance and then suddenly to cut off the current and withstand the resulting voltage which built up across the circuit. A schematic of this type of circuit is shown in Fig. 9. In this circuit the tube is required to pass about two amperes peak plate current which builds up over a period of about 150 microseconds. The d-c. voltage under these conditions is about 1000 to 4000 volts and the screen voltage may be obtained from the same source through series resistance. The grid is driven only slightly positive. Screen-grid dissipation is one of the limiting factors in this type operation. Primary emission from the screen

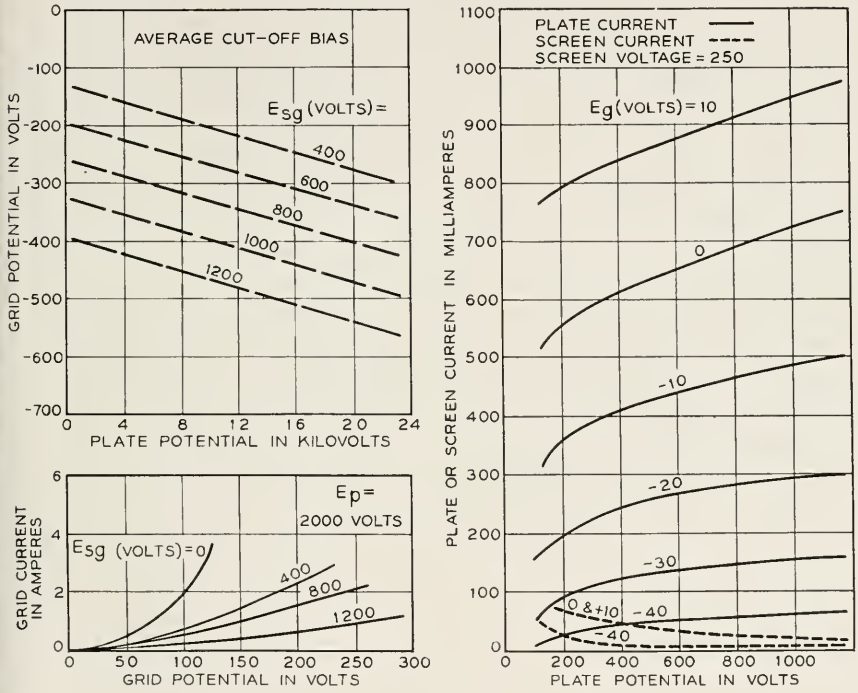


Fig. 11 (Continued)

grid, being present when the plate rises to its very high potential, tends to discharge the circuit prematurely, the energy wasted appearing as heat at the plate.

The 426XQ Tube

Since there was considerable demand for tubes capable of operating at voltages as high as 25 KV, a tube was developed to operate at this voltage. The limit of the 5D21–715B type structure seemed to have been reached at about 20 KV. It was also desirable to increase the current rating of the

tube since in a laboratory test equipment a pulse power of 1.5 to 2 megawatts was needed. The laboratory number 426XQ tube shown in Fig. 10 was the result. Four of these tubes in parallel were capable of providing pulses of about 1.75 megawatts. In the 426XQ tube, the bulb used on the 701A was employed and the plate supported entirely from its terminal in this bulb. The same four cathodes were used, but were spaced farther apart than in the 715-type tube. Two separate control-grids and two screen-grids were used, each pair encompassing two cathodes. This allowed a reduction of dissipation per grid compared to the 715-type, otherwise similar techniques were employed. The characteristics are shown in Fig. 11.

The tentative ratings applied to the 426XQ are given in Table I. The allowable peak plate current was increased for this tube because the technique of processing had improved so that a higher level of cathode activity was consistently realized. Also the greater spacing between cathodes and use of two sets of grids resulted in better grid cooling. The tube was not used in any radar equipment, because by the time it was available the trend in radar equipment was toward small, compact apparatus in which spark gap and transmission line modulators[5] found considerable application. The 426XQ proved very satisfactory in laboratory test equipment. One set of these tubes operated for somewhat more than 2000 hours.

The Chief Problems

The difficulties experienced with this series of oxide-cathode pulse modulator tubes can be divided into three general classifications, namely: sparking, cathode emission, and grid emission.

The sparking in these tubes can roughly be divided into two types, which may be called inter-electrode sparking and cathode sparking. Inter-electrode sparking is a discharge between two electrodes of the tube caused by the momentary breakdown of the insulation between them or by a gas discharge. If the breakdown of insulation is caused by light deposited films, the resultant discharge usually causes removal of the film and cures the trouble automatically, provided no other damage is done to the tube. Gas discharges from isolated pockets may be initiated by the high fields or by bombardment by stray electrons. If these pockets are not numerous they are usually dissipated after a few minutes of tube operation such that further sparking is very intermittent and probably not of sufficient intensity to interfere with operation. The gas so released is ordinarily taken up by the getter in the tube so that operation is not subsequently impaired.

Cathode sparking may be caused by positive ion bombardment of the cathode or by poor adherence of cathode material when subject to electro-

static fields. This type of sparking usually does not clear up and when it becomes serious the tube must be replaced.[6] It can be aggravated by sparking in the oscillator part of the radar system. There is some evidence to indicate that very high rates of rise of the pulse current drawn from the cathode may tend to produce cathode sparking. At rates of rise in excess of about 50 amperes per microsecond per square centimeter of cathode area a tendency for increased sparking has been noticed.

Cathode emission, here as in any other tube, is governed by cathode temperature and other considerations such as quantity and kind of gas in the tube, the core material, coating material, and techniques of processing. No attempt will be made to consider these factors in this paper as they are sufficiently complex that no very clear cut dissertation can be given. Standard core materials and coatings were employed with good results. It was found that the double carbonates (Ba, Sr) were less subject to sparking than the triple carbonates (Ba, Sr, Ca). The cleanliness and previous treatment of the other parts of the tube seemed to be the major factor in determining the level of emission obtained.

Primary grid emission, or thermionic emission from the control-grid and screen-grid, was one of the most difficult problems in the development and production of these tubes. Many trials were made using different materials and coatings on the grids, but from all considerations gold was found to be the most satisfactory. The grids in all the tubes described here are gold plated or gold clad molybdenum. It is not considered that the use of molybdenum for the core material is necessary, it being used here mainly because it seemed to be the most economical material that had sufficient stiffness to maintain grid alignment. Materials that tend to alloy with gold easily are not suitable as it was found that gold alloys were not as good as pure gold on the grid surface. The limitation involved in the use of gold is that the temperature of the grid must be kept low enough that evaporation of gold is not serious. This temperature limit is probably about 700°C. If gold is evaporated, the grid soon loses its coating and primary emission builds up rapidly. Also, the cathode emission seems to be poisoned by the gold vapor.[7]

Acknowledgment

The author wishes particularly to acknowledge the contributions of his immediate associates, Messrs. H. L. Downing, J. W. West, and J. E. Wolfe in the development of this series of tubes. Many others also made important contributions. We are also indebted to the M.I.T. Radiation Laboratory for data from life tests which they conducted on many of these tubes.

References

1. R. Colton, Radar in the U. S. Army, *Proc. I.R.E.*, Vol. 33, pp. 740–750, November 1945.
2. The SCR-268 Radar, Editorial, *Electronics*, Vol. 18, pp. 100–109, Sept. 1945.
3. W. C. Tinus and W. H. C. Higgins, Early Fire-Control Radar for Naval Vessels, *Bell Sys. Tech. Jour.*, Vol. 25, pp. 1–47, Jan. 1946.
4. E. Peterson, Coil Pulsers for Radar, *Bell Sys. Tech. Jour.*, Vol. 15, pp. 603–615, Oct., 1946.
5. F. S. Goucher, J. R. Haynes, W. A. Depp, and E. J. Ryder, Spark Gap Switches for Radar, *Bell Sys. Tech. Jour.*, Vol. 15, pp. 563–602, October 1946.
6. E. A. Coomes, The Pulsed Properties of Oxide Cathodes, *Jour. Applied Phys.*, Vol. 17, pp. 647–654, August 1946.
7. J. Rothstein, The Poisoning of Oxide Cathodes by Gold, (Abstract) *Phys. Rev.* Vol. 69, 1st/15th June 1946, p. 693.

Polyrod Antennas

By G. E. MUELLER and W. A. TYRRELL

The polyrod, a new form of microwave endfire antenna, is described. This consists of a properly shaped dielectric rod protruding from a metal waveguide. For applications requiring moderate gain, it possesses desirable electrical and mechanical properties. It is useful as a unit antenna in broadside arrays on account of its low crosstalk into adjacent polyrods. This paper describes work done from 1941 to 1944 at the Bell Telephone Laboratories, Holmdel, N. J. Important individual contributions are acknowledged in some of the footnotes. A report of this development has been withheld from earlier publication for reasons of military security.

1. Introduction

A UNIFORM rod (or "wire") of dielectric material without metallic boundaries is a well-known type of single conductor transmission line. In this kind of waveguide, a portion of the energy travels along in the space outside the rod. At discontinuities, including those caused by proximity to other objects, radiation takes place. For this reason, the dielectric waveguide has not become generally useful as a transmission medium, this need having been satisfied by the hollow metal pipe. The tendency toward radiation inherent in the dielectric guide is turned to advantage, however, in a new form of radio antenna. Here the objective is to encourage radiation from all parts of the dielectric rod. In progressing along the rod, therefore, power is gradually transferred from within the dielectric to the space outside. At a point where the transfer has been effectively completed, the rod can be terminated abruptly. By proper design, this radiating structure is an endfire antenna. Since it has been most often fabricated from polystyrene, it has become known as the polyrod antenna. It is especially useful for microwaves.

We must now review and examine certain features of dielectric rod transmission and of endfire antenna theory, for their bearing on polyrod design and performance.

2. Dielectric Wire Transmission[1]

A dielectric rod can be energized with an infinite variety of transmission modes. These are in general hybrid waves[2] possessing transverse and longitudinal components of both E and H. We shall here be concerned only

[1] Hondros and Debye, *Ann. der Phys.*, Vol. 32, pp. 465–476; J. R. Carson, S. P. Mead and S. A. Schelkunoff, *B.S.T.J.*, Vol. 15, pp. 310–333, 1936; G. C. Southworth, *B.S.T.J.*, Vol. 15, pp. 284–309, 1936; S. A. Schelkunoff, "Electromagnetic Waves," pp. 425–428, D. Van Nostrand, New York, 1943.

[2] Except in the case of circular symmetry. Cf. Schelkunoff, *loc. cit.*, pp. 154, 425.

with the lowest mode,[3] that which is the counterpart of the dominant wave[4] in a metal pipe. If a dielectric-filled metal guide is excited at the dominant mode, and if the metal shield is abruptly terminated, the wave energy will continue on in the unsheathed dielectric rod and will be confined almost exclusively to the lowest hybrid mode. This is, indeed, the most common way of exciting the dielectric wire.

The extent to which the power is concentrated within the dielectric is a function of the rod diameter and dielectric constant. This is shown[5] in Fig. 1. If the curves for the two different dielectric constants are replotted against the effective diameter, $\frac{D\sqrt{\epsilon}}{\lambda}$, they become more nearly

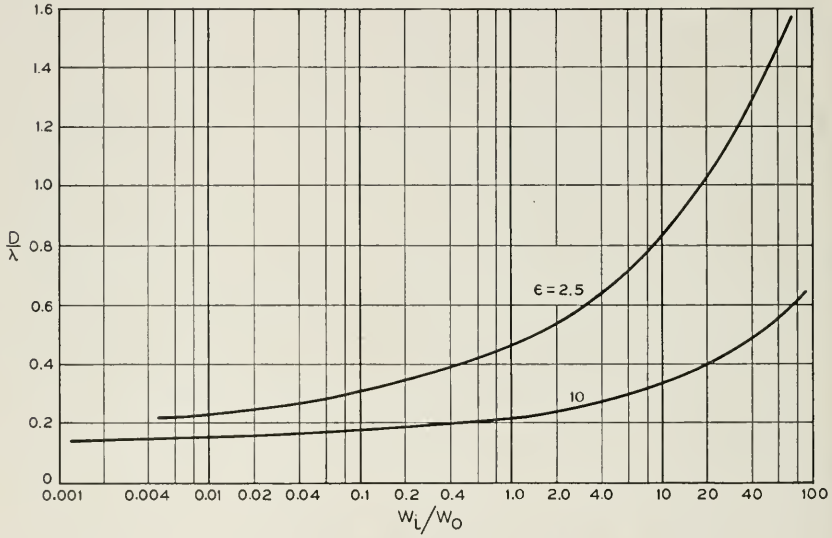


Fig. 1—Ratio of power inside W_i to power outside W_o for a cylindrical dielectric wire.

coincident. A universal curve cannot be given, however, because the field-retaining effect of a dielectric-air interface increases with increasing dielectric constant.

The phase velocity within the rod is also a function of the diameter and dielectric constant, as shown[5] by Fig. 2. When $\frac{D}{\lambda}$ is very small compared with unity, the rod exerts negligible guiding action, and the transmission is close to that in free space. For rods of large diameter, the power is con-

[3] Unlike all other modes in a dielectric wire and all modes in a conducting pipe, the lowest dielectric wire mode theoretically has its cutoff at zero frequency. Cf. Schelkunoff, *loc. cit.*, p. 428.

[4] That is, the TE_{11} mode in circular pipe or the TE_{10} mode in rectangular pipe.

[5] Figs. 1 and 2 are based on calculations by Dr. Marion C. Gray.

fined almost entirely within the rod, and the phase velocity approaches that in an unbounded medium of the same dielectric constant. By choosing intermediate values of $\frac{D}{\lambda}$, $\frac{v}{c}$ can be varied between these limits.

3. Endfire Antennas

We consider a linear array of isotropic radiators, infinite in number but so closely spaced as to occupy a finite length. We assume that the radiators are uniformly excited from a feed line, a transmission line parallel to the array phasing the various elements according to phase velocity on the line. The radiation pattern is given by[6]

$$r = \left| \frac{\sin \pi(\rho \cos \theta - \beta)}{\pi(\rho \cos \theta - \beta)} \right| \tag{1}$$

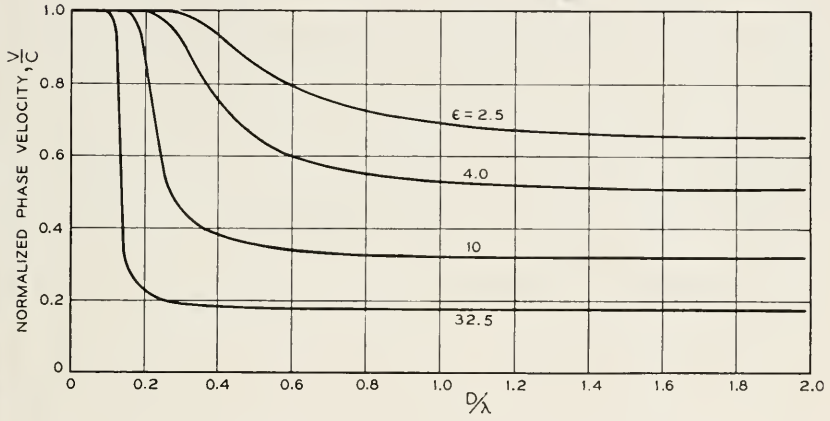


Fig. 2—Normalized phase velocity for a cylindrical dielectric wire.

where r = relative field strength
θ = angle with respect to the array axis
ρ = length of array in free space wavelengths
$2\pi\beta$ = phase shift in radians in the feed line from one end of array to the other end.

The pattern is symmetrical about the array axis.

Plotted from (1), Fig. 3 shows the pattern of a six wavelength radiator, $\rho = 6$, for selected values of β. When $\beta = \rho$ (= 6 in this case), phase velocity along the feed line is equal to free space velocity, and the resulting pattern is endfire. With $\beta = \rho + 0.5$ (= 6.5) the pattern remains endfire and the major lobe becomes sharper. For $\beta < \rho$ and $\beta \gtrsim \rho + 0.5$ (as shown by $\beta = 5.0, 5.5, 7.0$) the pattern deteriorates into a forward conical beam.

[6] R. M. Foster, *B. S. T. J.*, Vol. 5, p. 307, 1926.

The gain of a uniformly excited endfire antenna[7] with $\beta = \rho$ is 4ρ. For $\beta \neq \rho$, the gain can be written

$$g = 4A\rho. \tag{2}$$

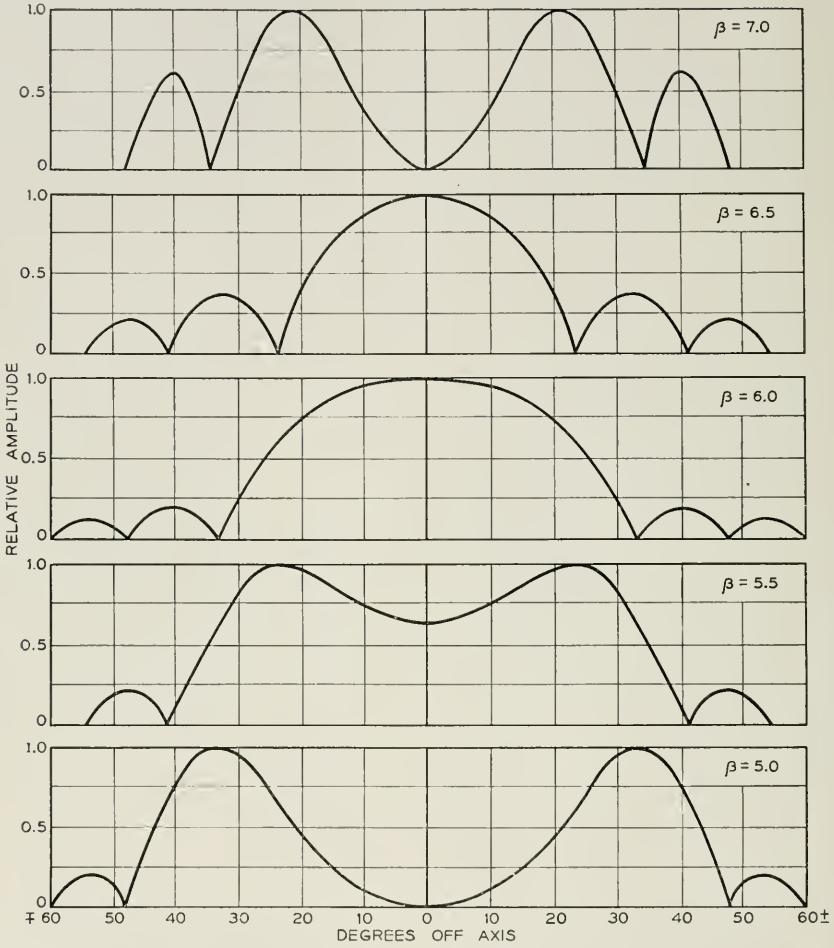


Fig. 3—Directional patterns of a six wavelength ($\rho = 6$) continuous array.

The factor A is given graphically in Fig. 4 as a function of $2\pi(\beta - \rho)$, the phase lag.[8] The highest gain occurs for a phase lag of approximately π radians relative to free space transmission, that is, for $\beta = \rho + 0.5$, in conformity with the patterns of Fig. 3. For a short radiator, A is about 2; with increasing antenna length, A approaches 1.8.

[7] Schelkunoff, *loc. cit.* p. 347.
[8] Fig. 4 and equation (5) were supplied by Dr. H. T. Friis.

The width of the major lobe is given by

$$\text{Beam Width} = \frac{B}{\sqrt{\rho}}. \tag{3}$$

The constant B depends on $\beta - \rho$ and on the manner in which beam width is defined. For width in degrees between half power points, and with $\beta - \rho = 0.5$ for maximum gain, B is computed from (1) to be about 60.

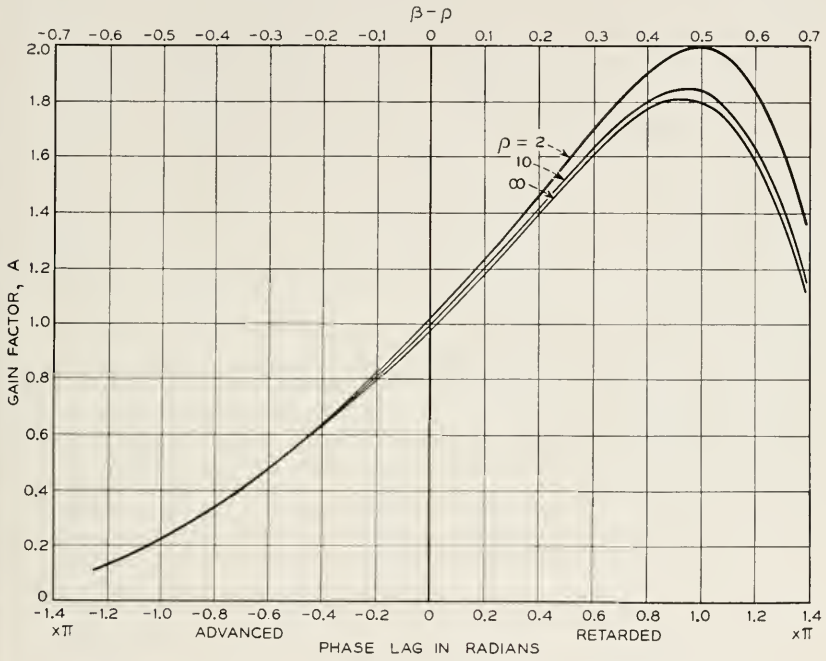


Fig. 4—Gain factor A as a function of phase lag in endfire arrays.

If a sinusoidal variation in excitation voltage along the radiator is superposed on the constant amplitude assumed for (1), we get

$$r = \left| a \frac{\sin \pi(\rho \cos \theta - \beta)}{\pi(\rho \cos \theta - \beta)} + (1 - a) \frac{\cos \pi(\rho \cos \theta - \beta)}{1 - 4(\rho \cos \theta - \beta)^2} \right| \tag{4}$$

where a is defined in Fig. 5. This figure gives patterns of a six wavelength radiator according to (4) for various values of a. Here β is fixed at 6.5 for maximum gain. Tapering symmetrically away from the center decreases the minor lobes. The gain is also decreased, but to a lesser extent.

Exponential tapering comes about from heat losses and radiation losses in the feed line. With attenuation α per wavelength, (1) becomes[8]

$$r = \sqrt{\frac{2 \cosh \alpha\rho - 2 \cos 2\pi(\cos\theta - \beta)}{\alpha^2\rho^2 + 4\pi^2(\rho \cos\theta - \beta)^2}} \tag{5}$$

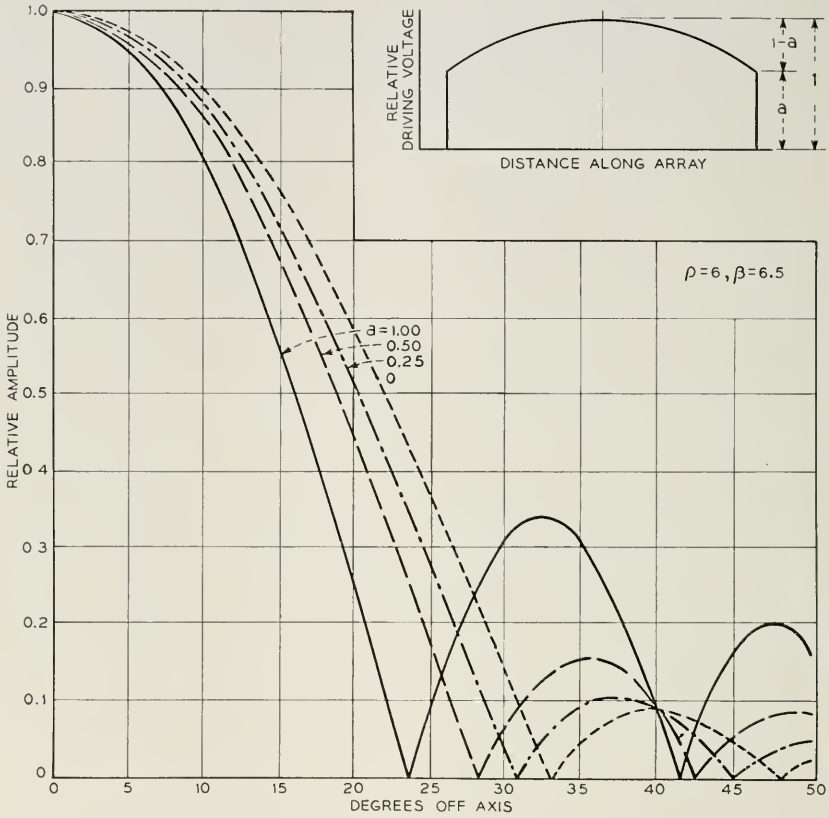


Fig. 5—Effect of sinusoidal tapering of power upon directional characteristic of a six wavelength continuous array.

Feed line attenuation increases slightly the minor lobe amplitudes and fills in the nulls. Exponential tapering caused by radiation can be reduced or eliminated if the coupling of the radiating elements to the feed line is gradually increased along the line.

4. The Polyrod Antenna

It has been found experimentally that a suitably proportioned dielectric rod can act as an efficient endfire radiator. A complete understanding of

[8] Loc. cit.

its operation involves the solution of Maxwell's equations subject to the boundary conditions appropriate to the configuration. An analysis of this sort is not available because of its mathematical complexity. However, a satisfactory explanation of polyrod operation, especially for engineering purposes, can be obtained by establishing analogies with array theory, coupled with existing knowledge about transmission in uniform dielectric wires. In this treatment by analogy, we remain essentially ignorant of the local fields in the vicinity of the dielectric, the role played by the discontinuities at both ends of the antenna, and other detailed features. We do have, however, a working theory which predicts closely the features of the radiation as observed at a distance. Under these circumstances, insistence upon a rigorous field solution has not so far appeared necessary.

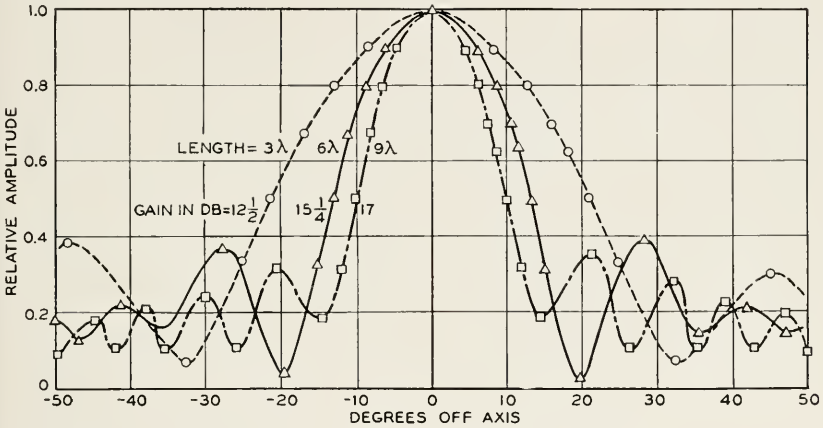


Fig. 6—Data on polyrods of uniform rectangular cross-section $\frac{1}{3}\lambda$ by $\frac{1}{2}\lambda$.

Experimental data have been obtained at frequencies in the vicinity of 3000 megacycles except for Fig. 9, representing work at 9000 megacycles. For the sake of generality, these results are presented in dimensions of λ, the free space wavelength. In all cases, polyrods have been energized from a dielectric filled metal guide whose conducting sheath is abruptly terminated, the dielectric continuing on as the radiator.

The earliest form of polyrod[9] was a polystyrene rod of uniform rectangular cross-section, about $\frac{1}{3}\lambda$ by $\frac{1}{2}\lambda$. Figure 6 shows the gains and directional patterns measured for such rods in three different lengths. In a plane normal to the axis, the radiation is approximately isotropic. The observed gains are proportional to length. They are greater than 4ρ by a factor of

[9] The earliest work on polyrods was done in 1941 by Dr. G. C. Southworth. Cf. his *U. S. Patent* 2,206,923 issued in 1940.

about 1.4. Phase velocity in these rods was not measured and is not available from theory. Referring to Fig. 4, however, we must assume at least 0.4π radians of phase retardation to explain the increased gain. When the pattern for the 6λ rod is compared with the sharpest pattern ($\beta = 6.5$) in Fig. 3, the observed characteristic is sharper than expected even with a phase retardation of π. The amplitudes of minor lobes are in good agreement. Attenuation, as revealed by the amplitudes at minima in the patterns, is apparently appreciable but not serious.

The principal defect of the uniform polyrod is the strong minor lobes. This is remedied by tapering the amplitude of radiation symmetrically about the midpoint, as suggested in Fig. 5. To obtain such tapering let us start at the waveguide end with a relatively thick rod. From Fig. 1, this tends to retain a larger fraction of the power and should therefore not radiate so strongly. Let us decrease the cross-section gradually in progressing along the rod, thus increasing the power radiated. Upon reaching a point near the center, we find the power in the rod already considerably diminished by the radiation which has already taken place. Beyond this point, gradually decreasing radiation is automatically secured with a uniform cross-section as a result of previous radiation.

This line of reasoning, calling for a polyrod tapered down in cross-section only in the first half of its length, is verified experimentally. Since detailed field analysis is not available for the polyrod, the most favorable proportions have been found empirically. Three examples will be described.

Figure 7 shows a 6λ rectangular polyrod linearly tapered for a little more than half its length from a base $\frac{1}{2}\lambda$ square to a rectangular section $\frac{1}{4}\lambda$ by $\frac{1}{2}\lambda$. the remainder being uniform. The tapering is confined to the magnetic plane. Measured phase velocity and directional pattern are included in Fig. 7. By reference to Fig. 5, the observed minor lobe amplitudes correspond to a value of a somewhat less than 0.5. The gain, considerably improved over the uniform rod, implies from (2) a value of 1.86 for A in remarkable agreement with Fig. 4.

Figure 8 shows data on a 6λ cylindrical polyrod linearly tapered for about half its length from a diameter of 0.5λ to 0.3λ with the remainder uniform. The pattern is very similar to that of the preceding example; the gain is slightly reduced, and $A = 1.66$. From Fig. 1, $\epsilon = 2.5$, about half the power is internal for $\frac{D}{\lambda} = 0.5$, while less than one-tenth is internal for 0.3. Agreement between Figs. 2 and 8 for phase velocity is fairly good.

Figure 9 gives information[10] about an 8.65λ radiator which resembles the

[10] Supplied by Mr. C. B. H. Feldman.

conical-cylindrical design of Fig. 8, but which is longer and is tapered for slightly less than half its length. The minor lobes (solid curve) are all lower than 0.125, a marked improvement over Fig. 8. From the measured gain, A is 1.82.

Regardless of whether the cross-section is square, rectangular, or round, radiation is nearly isotropic about the axis of the polyrod. For the patterns

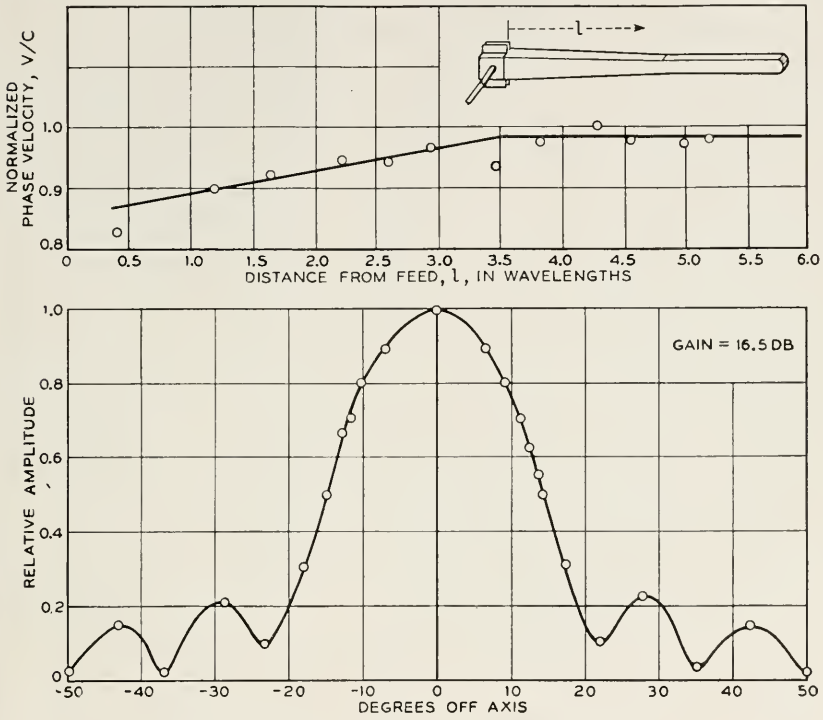


Fig. 7—Data on a 6λ tapered rectangular polyrod.

in Fig. 6–9, beam widths in degrees between half-power points correspond to values of B in (3) between about 50 and 60.

The characteristics of polyrods can thus be correlated with array theory for isotropic radiators continuously distributed along an axis. There are, to be sure, minor discrepancies which might become more serious in a different range of polyrod proportions. For the lengths and cross-sections tested, however, equations (1) to (5) describe polyrod performance very satisfactorily for engineering purposes.

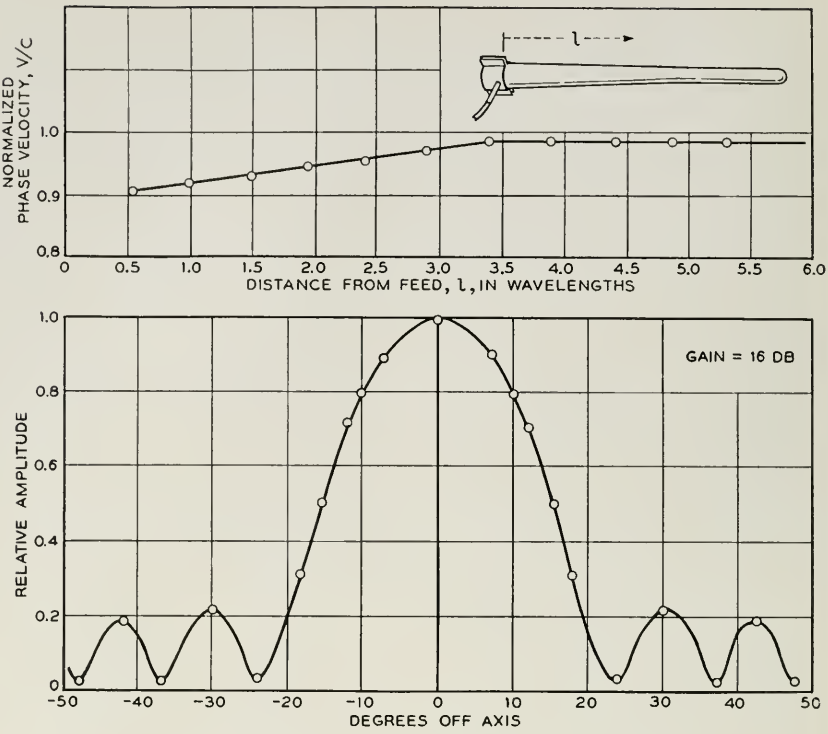


Fig. 8—Data on a 6λ tapered cylindrical polyrod.

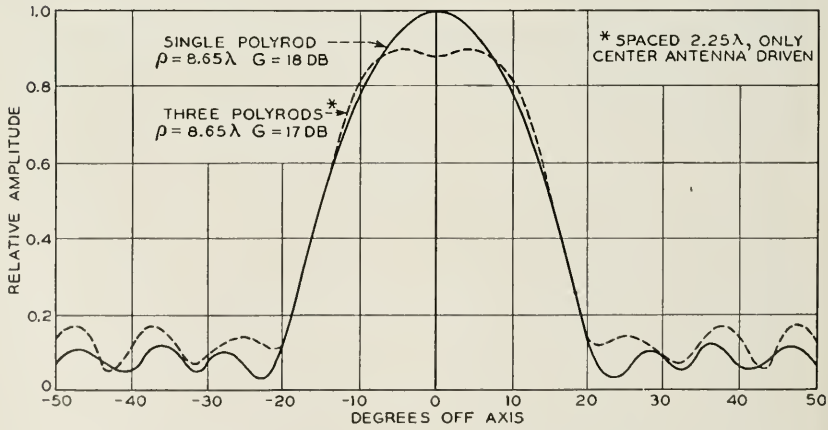


Fig. 9—Data on an 8.65λ tapered cylindrical polyrod, including effect of adjacent similar polyrods.

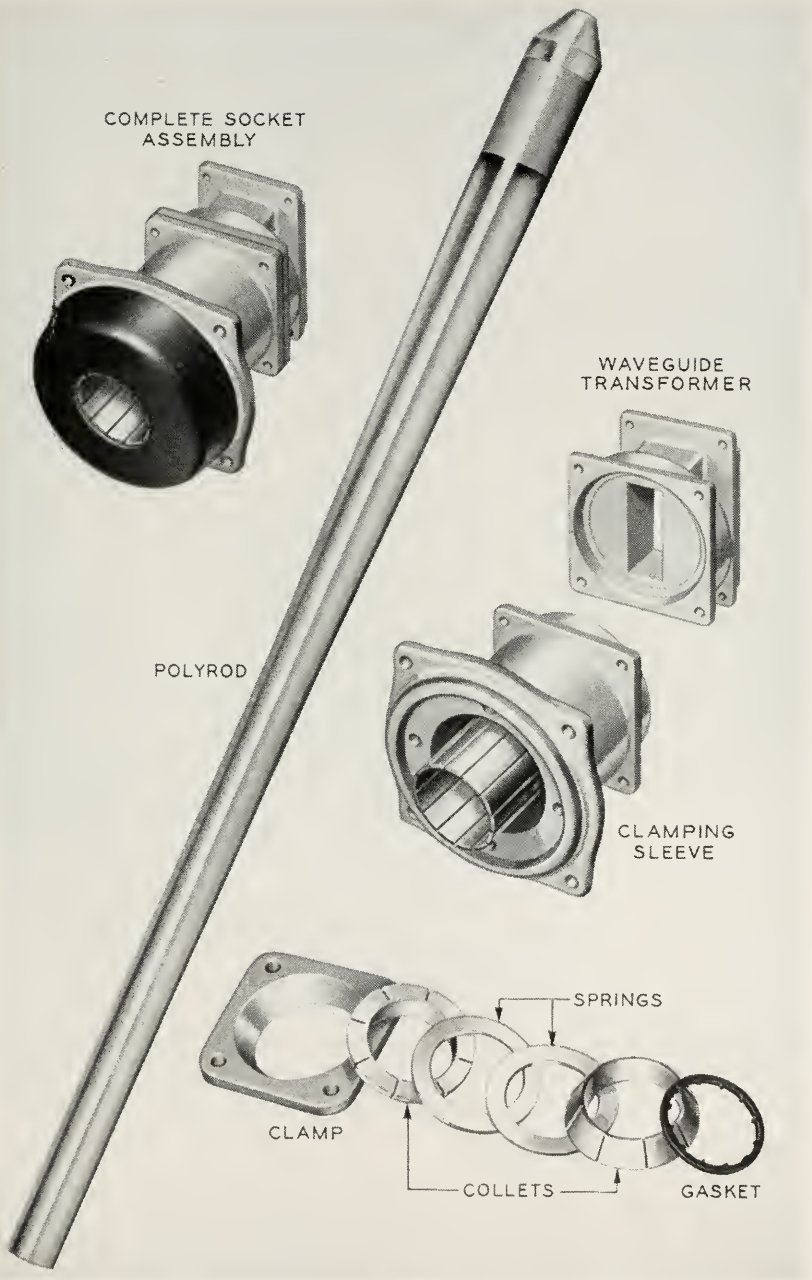


Fig. 10—Polyrod and waveguide feed details.

5. Construction and Operational Details

Figure 10 shows a production model of a polyrod for 3000 megacycles, with means for matching it to a rectangular waveguide.[11] A two-iris transformer is used with a resulting width of 4% between the 1 *db* standing wave points. The clamping illustrated is designed to maintain a firm grip on the rod despite tendencies of the polystyrene to cold flow.

Another type of coupling is indicated in Fig. 11. Here the polyrod is still fed from a waveguide but this is in turn transformed to a coaxial line. The composite can thus be regarded as a coaxial to polyrod coupling. The coaxial line taps at point *b* onto the short-circuited antenna *a-b-c* at a point

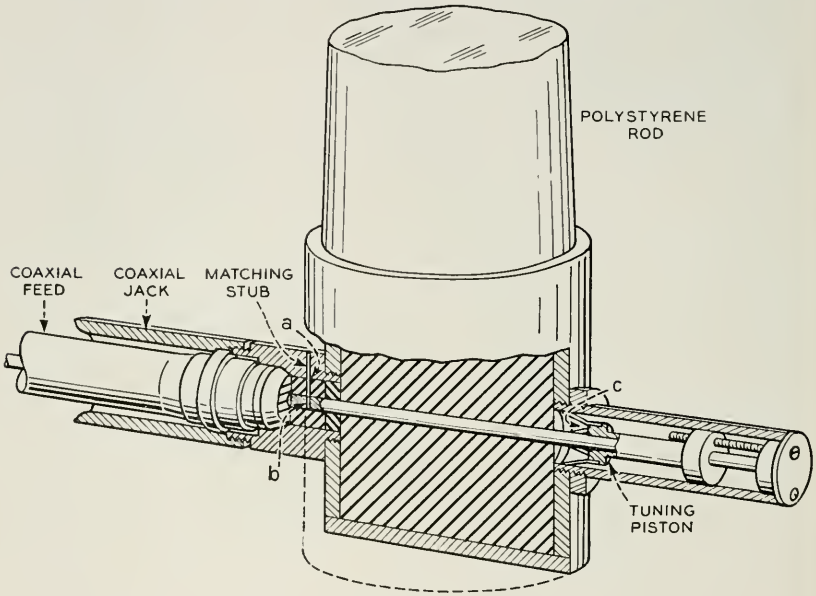


Fig. 11—Coaxial feed for polyrod.

chosen to match the characteristic impedance of the coaxial line. The back end of the waveguide is short circuited by a metal cap a quarter wavelength behind the transverse wire antenna. A movable coaxial plunger provides tuning. This arrangement has a bandwidth of 1% to the 1 *db* standing wave points.

The frequency response of a polyrod is inherently broad. The directive pattern varies slowly with both phase velocity and amplitude distribution along the axis. As shown in Figs. 1 and 2, these quantities are slowly varying functions of λ over a considerable range of polyrod proportions. At

[11] Developed by Mr. D. H. Ring.

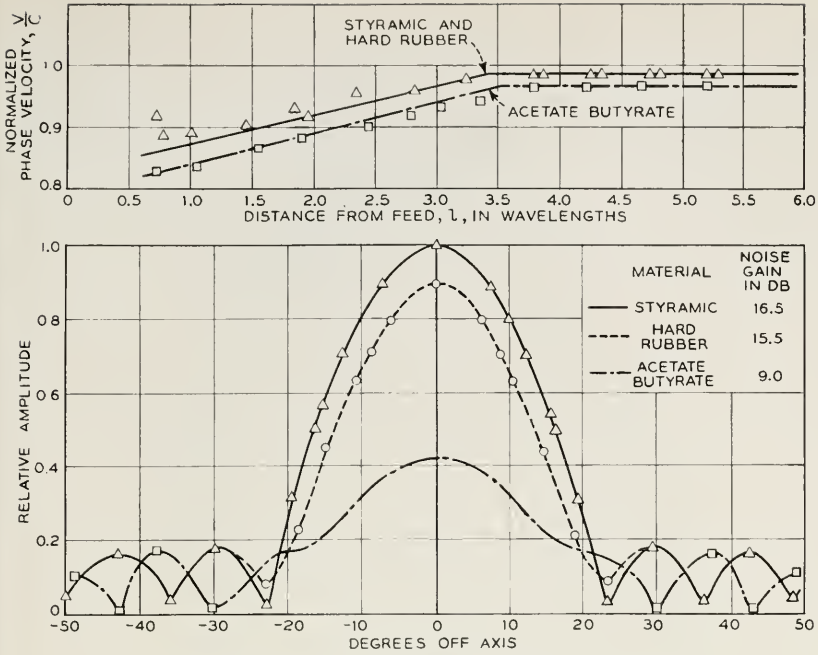


Fig. 12—Effect of dielectric loss on polyrod performance.

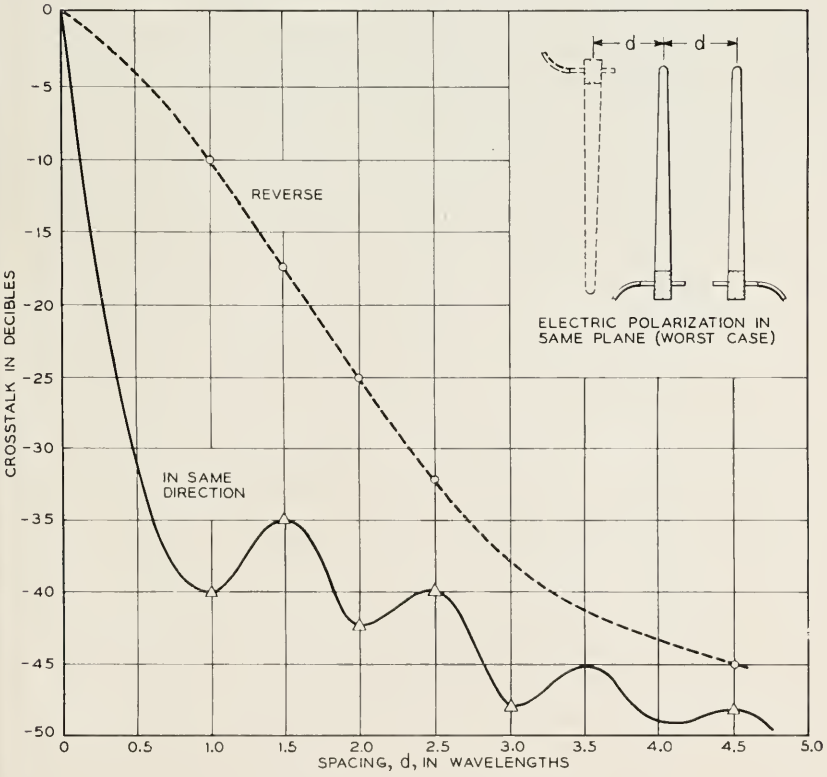


Fig. 13—Crosstalk between polyrods.

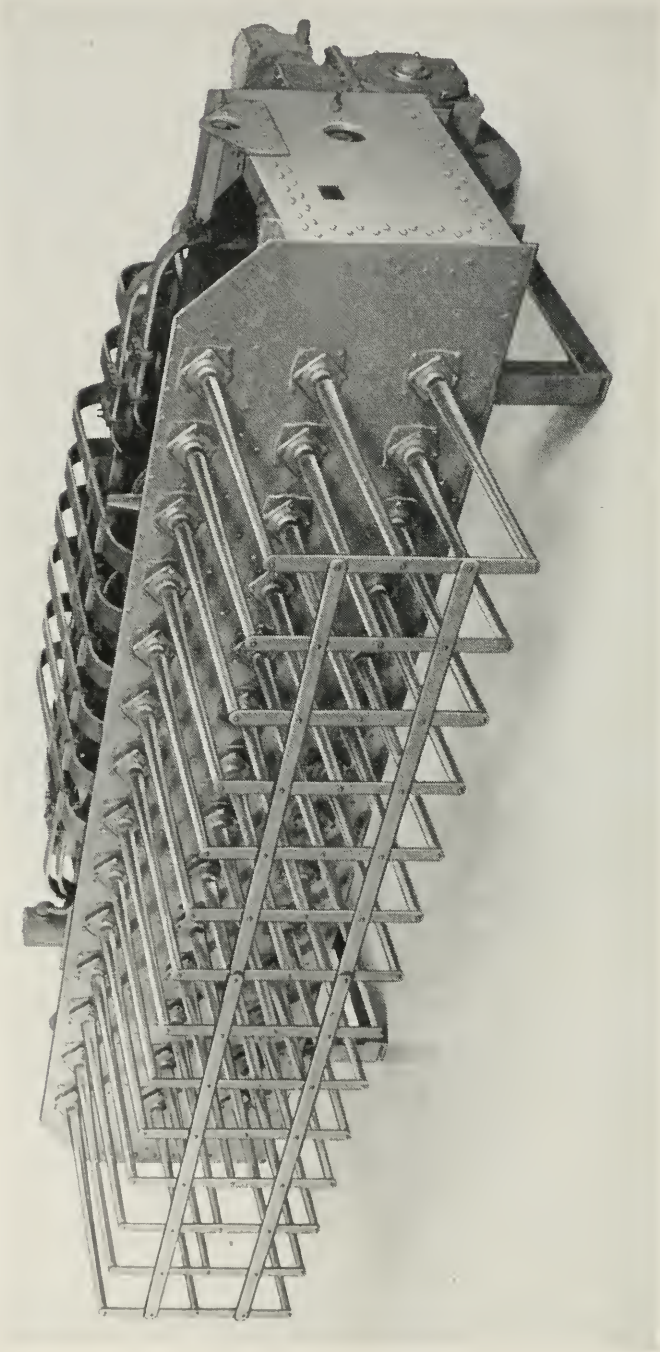


Fig. 14—Broadside array of polyrods.

present, the usable bandwidth is therefore limited primarily by the frequency response of the coupling arrangements from polyrod to waveguide or coaxial line.

We have been exclusively concerned so far with plane polarized radiation. A circularly symmetrical polyrod such as in Fig. 8 can be used equally well to radiate circularly polarized waves. To do this, the polyrod is fed from a waveguide in which circulary polarized dominant waves are generated by means of a 90° phase shift section.[12]

The effect of dielectric loss upon polyrod performance is shown in Fig. 12, to be compared with Fig. 7. The power factors are: Styramic, 0.0005; hard rubber, 0.003; acetate butyrate, 0.020; polystyrene, 0.0002. Materials having power factors less than 0.001 are satisfactory for polyrod antennas.

Figure 13 shows the crosstalk between adjacent polyrods, that is, the power received in one radiator when the other is energized. For polyrods pointing in the same direction, separations greater than a wavelength insure low mutual coupling. This makes the polyrod attractive as the element in broadside arrays. Proximity to other undriven polyrods affects the gain and directional pattern to a greater extent, as shown in Fig. 9.

More generally, the performance of a polyrod is affected by proximity to any metallic or dielectric objects. The gain and pattern must be determined empirically for each new configuration. It has been found that a metal rod can be placed parallel to a polyrod without seriously affecting its behavior so long as a separation of a wavelength or more is maintained. Sheets of dielectric material can be brought even closer without adverse effect so long as large surfaces are not in direct contact with the polyrod. These and other experiences suggest that the polyrod is relatively unaffected by nearby objects.

Tests have been made of the effect of fresh and salt water in the form of a spray or solid stream playing on a polyrod. Provided that puddles do not form on the surface, as can happen with rectangular polyrods, the effect is a decrease of 1 to 2 *db* in gain under the worst conditions.

In conclusion, for microwave applications involving moderate gains of 15 to 20 *db*, the polyrod assumes a convenient physical form and displays high electrical efficiency. It is less subject to disturbance by nearby objects than might be expected. It is especially useful as an element in broadside arrays. As an example of such arrays, Fig. 14 shows a 42 rod steerable beam antenna used in an important type of Navy fire control radar.

[12] For a discussion of this subject, cf. A. G. Fox, "An Adjustable Waveguide Phase Changer," to be published in *Proc. I. R. E.*

Targets for Microwave Radar Navigation

By SLOAN D. ROBERTSON

The effective echoing areas of certain radar targets can be calculated by the methods of geometrical optics. Other more complicated structures have been investigated experimentally. This paper considers a number of targets of practical interest with particular emphasis on trihedral and biconical corner reflectors. The possibility is indicated of using especially designed targets of high efficiency as aids to radar navigation.

Introduction

IT NOW seems likely that radar, developed during the war, will find increasing application as a navigational aid for aircraft and surface vessels. In fact there are good reasons for expecting that peace-time radar can be made even more efficient than its war-time prototype.

There are two ways of improving radar performance. One may concentrate on the radar set proper with the object of increasing either the power of the transmitter or the sensitivity of the receiver. Or, one may take steps to improve the echoing efficiency of the targets. The latter was, of course, not possible during the war since most of the targets of interest were controlled by the enemy. It is a purpose of the present paper to consider the design of various targets of high echoing efficiency and wide angular response which may be placed at strategic points as aids to radar navigation. The ideal reflector to serve as a "beacon" or "buoy" for guiding radar-equipped aircraft or ships would present a highly effective area to incident radiation over a full 360° in azimuth, and would also be operative over a fairly broad vertical angle. The value of a particular target for navigational purposes may therefore be considered in terms of two factors: effective area, and angular response.

The echo received by a radar from a particular target can be calculated by the formula:[1]

$$W_R = W_T \frac{A_R^2 A_{eff}^2}{\lambda^4 d^4} \tag{1}$$

where W_R = echo power available at the terminals of the radar antenna.
W_T = power launched by radar.
A_R = effective area of radar antenna assuming that the same antenna is used both for transmission and reception.

[1] This equation follows directly from Equation (1) of a paper by H. T. Friis, "A Note on a Simple Transmission Formula," *Proc. I.R.E.*, Vol. 34, pp. 254–256, May 1946. The radar transmission formula is obtained by applying Friis' formula twice; first to the transmission from the radar to the target, then to the transmission from the target to the radar.

A_{eff} = effective area of target.[2]
λ = wavelength.
d = distance between radar and target.

The above formula applies to the case where free-space propagation prevails; that is, where multiple path or anomalous transmission effects are absent.

It is apparent from the formula that, at a given wavelength and range, the received echo power can be increased by increasing the transmitted power, the size of the antenna, or the effective area of the target. The present paper will consider only the latter.

In some cases the effective area of a target can be calculated from simple geometrical optics. For the more complicated structures it is always possible to measure the effective area by comparing the signal reflected by the object in question to the signal reflected from a simple target of known effective area.

Flat Plates

The simplest target for which the effective area can be calculated is a flat metal plate oriented so as to be perpendicular to the incident radiation. It can be demonstrated that a flat plate with all linear dimensions large in proportion to the wavelength of the incident radiation has an effective area which is substantially equal to its geometrical area. Diffraction effects at the edges of such a plate are small in comparison with the energy reflected from the central portion of the plate.

Flat plates, however, have the serious disadvantage that, in order to create strong echoes, they must be maintained accurately perpendicular to the incident rays. At other angles of incidence the echoes fall off rapidly. For this reason flat plates are of limited value as targets for use in navigation.

Dihedral Corner Reflectors

A dihedral corner reflector consists of two perpendicular, plane conducting surfaces which are usually arranged so that they intersect along a common line. Figure 1 shows a typical dihedral reflector. The dihedral reflector has the important property that a ray which enters the corner will experience a reflection from each of the surfaces and will return in the direction from which it came, provided of course that the entering ray lies in a plane which is perpendicular to the line of intersection of the planes which form the

[2] The term "effective area" as used in this paper refers to the equivalent flat plate area of a target. The echoing effectiveness of a target may alternatively be expressed in terms of the cross section of an equivalent isotropic reflector as described by Schneider, "Radar," *Proc. I.R.E.*, Vol. 34, p. 529, August 1946. The alternative unit is called the "scattering cross section" and is frequently denoted by the symbol σ, although Schneider uses S. The two quantities are related by the equation $\sigma = 4\pi A^2_{eff}/\lambda^2$. Both units are useful. For most of the targets considered in the present paper, A_{eff} does not vary with λ and is therefore preferable.

corner. The latter restriction constitutes the principal objection to the practical use of dihedrals. The path of a typical ray is shown in Fig. 1.

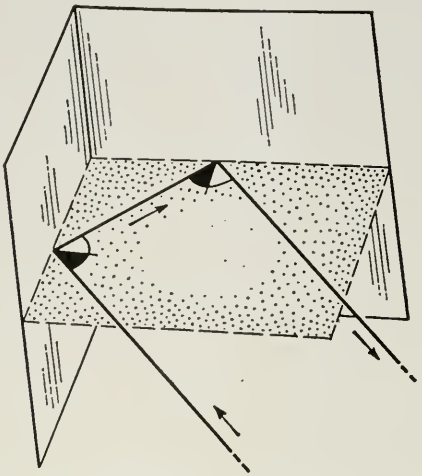


Fig. 1—Dihedral corner reflector.

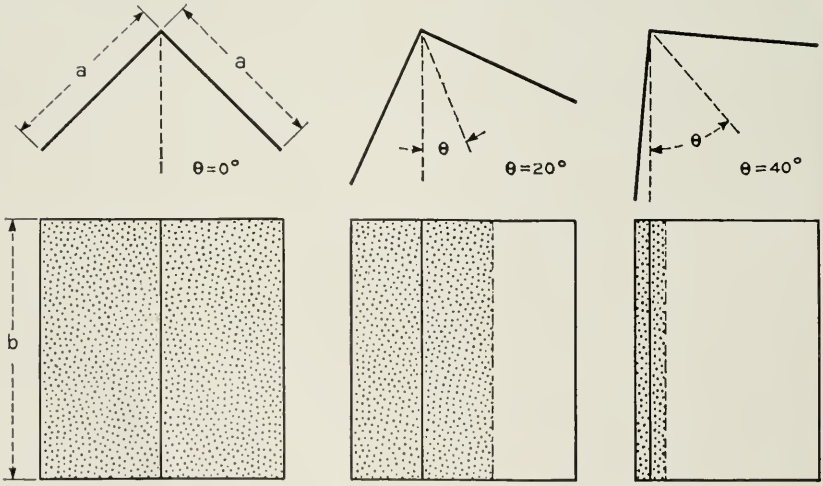


Fig. 2—Variation of effective area of a dihedral with aspect angle.

The effective area of a dihedral reflector depends upon both the size of the reflector and the orientation of the reflector with respect to the incident rays. Figure 2 shows how the effective area varies as the dihedral is rotated about the line of intersection of the two planes. The effective areas for the different orientations are shown by the shaded regions in the lower part of the

figure. For a reflector having the dimensions shown in the figure the effective area for different angles of incidence θ can be calculated by the formula.

$$A_{eff} = 2\ a\ b \sin (45° - \theta)$$

where θ is always considered positive and less than 45°.

Figure 3 shows the polarization of the reflected ray for differently polarized incident rays. For our purpose, the incident rays may be assumed to enter the left side of the reflector shown in the figure and the reflected rays may be assumed to emerge from the right. It is apparent that if the incident ray is polarized either parallel or perpendicular to the line of intersection of the two surfaces the reflected ray will be polarized in the same plane as the incident ray. If the incident ray is polarized at an angle of 45° to the line of intersection, the reflected ray will be polarized perpendicularly to the incident ray. In the latter case the signal received back at the radar will not ordinarily be accepted by the same antenna which launched the incident radiation.

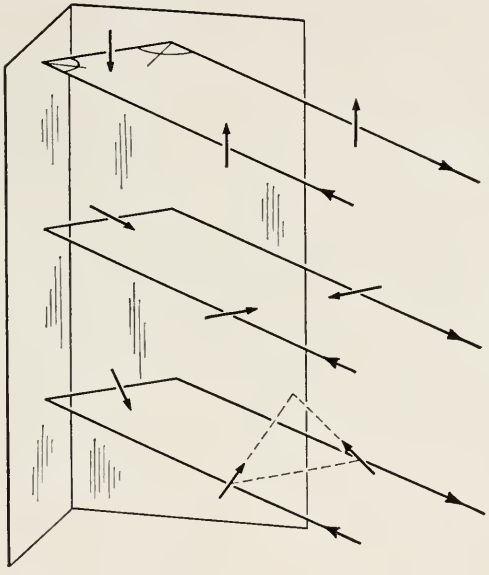


Fig. 3—Polarization effect in a dihedral reflector.

Trihedral Corner Reflector

Assume that three reflecting surfaces AOB, AOC, and BOC are placed so as to form the right-angled corner illustrated in Fig. 4. In general, electromagnetic waves, upon striking an interior surface of the device, will undergo a reflection from each of the three planes and return in a direction parallel to

and with the same polarization as the incident ray. The path of a typical ray is shown by line 1, 2, 3, the particular ray chosen having entered the reflector along a line perpendicular to the plane of the paper. Points 1 and 3 represent the initial and final points of reflection, respectively, whereas point 2 represents the intermediate reflection point.

Two important conclusions can be drawn from a careful inspection of the path 1, 2, 3; namely, the projections of points 1 and 3 are diametrically opposite on a circle drawn about point O as a center, and points 1 and 3 appear to be images of point 2; i.e., the ingoing ray at 1 appears to be directed toward

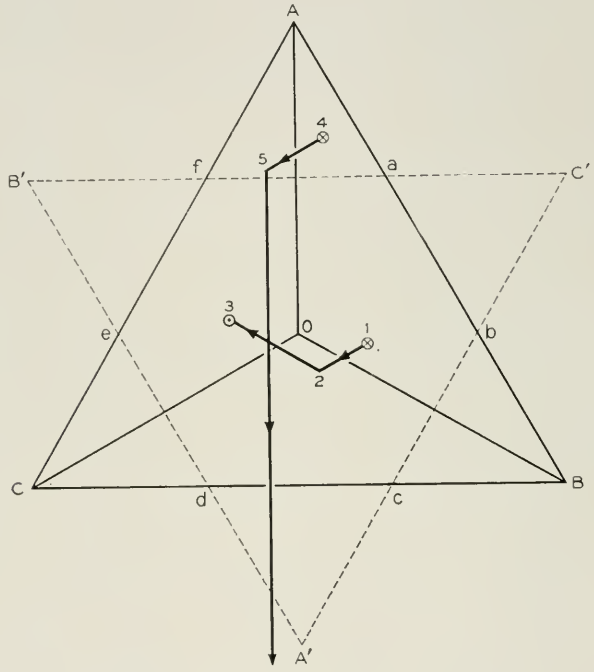


Fig. 4—Trihedral corner reflector showing the paths of typical rays.

the image of point 2 in plane AOB, and the outcoming ray at 3 appears to come from the image of point 2 in plane AOC.

Not all rays falling upon a corner reflector of finite dimensions will be reflected in the direction of the source. For example, a ray striking point 4 in Fig. 4 may be reflected successively at points 4 and 5, but if the plane BOC is not sufficiently extended it will not undergo the necessary third reflection required to return the ray in the incident direction.

The portion of the projected cross-section of a corner reflector which is able to return incident radiation to the source is called the "effective area." It is, of course, a function of the aspect, that is to say, the angle at which the

reflector is being viewed, as well as the geometrical configuration of the reflector. For some of the simpler configurations the effective area can be readily determined by the following procedure.

Project the aperture of the reflector through the apex O to form the image (A′ B′ C′ of Fig. 4); then project the aperture and its image upon a plane perpendicular to the incident rays. The area common to the projections of the aperture and its image is equal to the effective area. The effective area of the triangular reflector of Fig. 4 is, therefore, represented by the hexagon a b c d e f. Only those rays, perpendicular to the plane of the paper, which

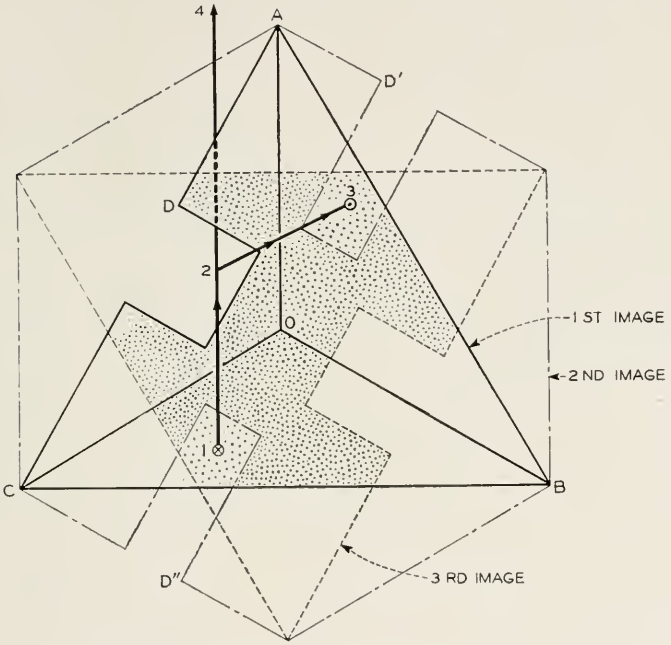


Fig. 5—Determination of effective area of trihedral corner reflector.

fall inside the hexagon will be returned. Exactly the same procedure is used in determining the effective area for other aspect angles.

The above rule must, however, be applied with caution. Situations arise in which rays falling upon the area determined by this method do not return to the source. Figure 5 shows a reflector in which this difficulty is encountered. This reflector differs from the previous reflector in that it has a notch cut in one of the reflecting surfaces. The projection of the aperture upon the plane of the paper is indicated by the solid line; that of its image by the dotted line. According to the rule of the preceding paragraph, one would expect the effective area to be defined by the total shaded area of the figure.

Such, however, is not the case. It was stated earlier that the ingoing rays appear to be directed toward the images of the intermediate reflecting points. This requires that the images of the intermediate reflecting points fall inside of the effective area. In Fig. 5, the images of the notch fall inside of what would otherwise be the effective area. Since the notch is incapable of serving as an intermediate reflector, the more lightly shaded areas are excluded from the effective area. In the absence of the notch, a ray entering at 1 would be reflected at 2 and emerge at 3. In the presence of the notch, however, it passes through plane AOC and escapes in the direction of 4.

Therefore, in order to determine the effective area of a corner reflector of arbitrary shape and aspect, one must take account of three loci of points defined by the aperture as follows:

1) The aperture itself

2) The locus of points determined by taking the direct mirror image of each point of the aperture with respect to each of the two surfaces of the trihedral not containing the point. For example, point D of Fig. 5 will have the images D′ and D″ with reference to planes AOB and BOC, respectively. The complete locus of points determined in this way is represented by the dot-dash line of Fig. 5.

3) Locus of points on aperture after each has been assumed to have been projected through the vertex. This image is pictured by the dotted lines of Fig. 5.

These three images of the aperture can, for simplicity, be referred to as the first, second, and third images, respectively. The effective area is the area common to the projections of the first, second, and third images of the aperture upon a plane passing through the apex of the reflector and perpendicular to the incident rays. For a given aperture and aspect, a corner reflector can theoretically be replaced by a flat plate located at the apex. The size and shape of the flat plate will vary with the aspect as well as with the configuration of the aperture. The above procedure has been of considerable aid in studying reflectors having apertures of arbitrary shape.

Although the graphical analysis just given is sufficient to enable one to compute the effective area of a reflector for any aspect angle, it is frequently more convenient to determine the complete response pattern of a reflector experimentally. Most of the experimental results reported in this paper were obtained with a 1.25 centimeter radar arranged as shown in Fig. 6. Echo levels were measured on the screen of a type-A indicator using a calibrated intermediate frequency attenuator to restore the signal to an arbitrary reference level. It is believed that the levels measured in this way are accurate to within $\pm \frac{1}{2}$ decibel. The coordinate system used in recording and presenting the data is given in Fig. 7. The reflector was mounted on a turntable which could be rotated about horizontal and vertical axes.

Curves for the response patterns of a corner reflector of triangular aperture are shown in Fig. 8. These curves were obtained with a reflector constructed of silver-painted plywood whose aperture was in the form of a 24-inch equilateral triangle. It had been previously determined that, with suitable paints, reflectors of this construction behaved exactly as though they were made of sheet metal.

Depending upon its angle of arrival, a ray may be reflected by a corner reflector in one of four ways. If the angle is too oblique, the ray may not be returned in the direction of the source at all. If the incoming ray is

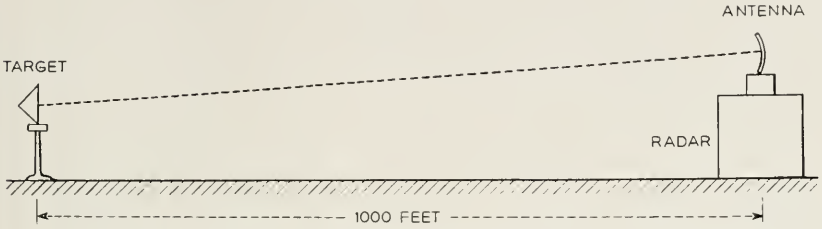


Fig. 6—Arrangement of apparatus for measuring effective areas of targets.

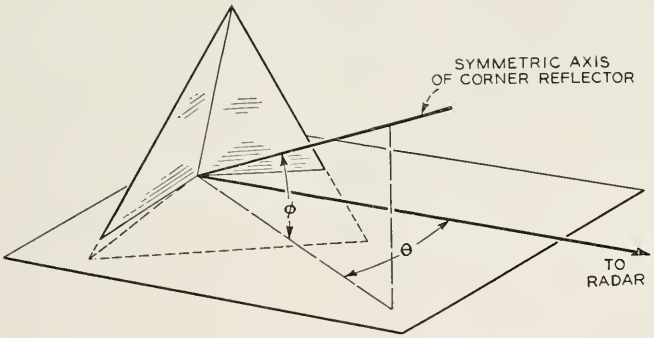


Fig. 7—Coordinate system used in presenting data.

exactly perpendicular to one of the three reflecting planes, it will be returned to its source after only one reflection. Should the ray arrive in a direction exactly parallel to one of the three planes, it will again be returned in the direction of its source but in this case it is reflected twice as in a dihedral. This particular mode of reflection is illustrated by the sharp peaks at the extremities of the curves in Fig. 8. For all remaining angles of approach the ray will be returned after three reflections in the manner already described. The central regions of the curves represent this type of reflection which is of principal interest in practice.

The effective area of the triangular trihedral reflector along the symmetric axis ($\theta = 0, \phi = 0$) can be computed from the geometry of Fig. 4.

$$A_{eff} = 0.289\,\ell^2 \tag{3}$$

where ℓ is the length of one side of the aperture such as CB. The effective area at other aspect angles can be computed by relating the echo level at the aspect in question with that along the symmetric axis.

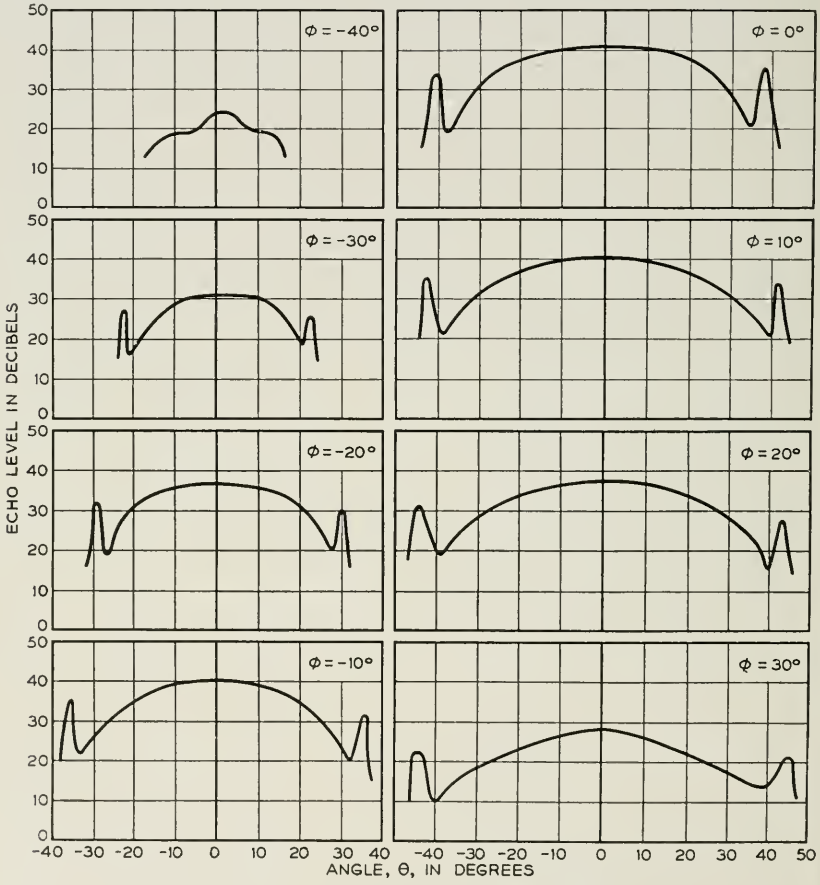


Fig. 8—Echo-response patterns of a triangular trihedral reflector.

It should be pointed out that the effective area for trihedral reflections is independent of wavelength where the reflector is large enough so that geometrical optics prevail. If the wavelength is increased, however, the sharp dihedral peaks at the edges of the pattern will be broader.

In the case of the triangular corner reflector the response levels for aspect angles of 30°, as measured from the symmetric axis, are down by 10 decibels. For many applications a flatter response pattern is desirable.

The present investigation led to the discovery that the response pattern of a corner reflector can be modified by a suitable alteration of the geometrical configuration of the aperture. There is even the suggestion that the response can, to a certain degree, be made to conform to a somewhat arbitrary pattern within a region extending to approximately 30° from the principal axis. The procedure for accomplishing this has, so far, been one of trial and error since the difficulties of a general mathematical solution appear to be

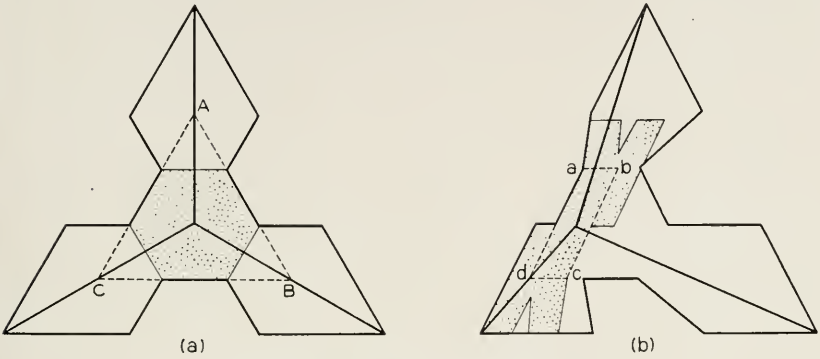


Fig. 9—Compensated corner reflector.

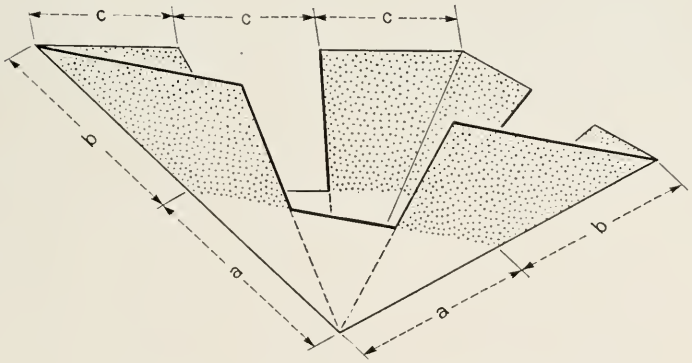


Fig. 10—Dimensions of face of compensated reflector.

insurmountable, at least in a practical sense. For practical purposes, however, it is comparatively easy to conduct a few graphical experiments in order to design a reflector having the desired response pattern.

Figures 9a and 9b show two views of a modified corner reflector which was designed to have a relatively flat response characteristic out to angles of 30° from the central axis. Each of the three sides of the reflector, instead of being triangular as formerly, has the contour shown in Fig. 10. The shaded regions of Fig. 10 represent the surface which has been added.

In Fig. 9a, one is assumed to be looking into the reflector along the symmetric axis. The effective area is represented by the shaded hexagon. Evidently, the effective area of the modified reflector is identical to the effective area of the original triangular reflector A B C. Therefore, for this particular aspect, the effective area has not been changed by the addition of the material at the corners.

Figure 9b is a view of the reflector at $\theta = 30°, \phi = 0°$. Again, the shaded region represents the effective area, and the parallelogram abcd is the effective area of the reflector before modification. The modification has evidently introduced a substantial gain in effective area for this aspect. A graphical comparison of the effective areas of Figs. 9a and 9b shows them to be of comparable magnitude.

With the dimensions defined as in Fig. 10, a corner reflector was constructed with a= b= 17″. The response curves of this reflector are plotted in Fig. 11, along with the curves of the ordinary triangular reflector. A substantial improvement in response is exhibited by the compensated reflector. In the region extending out to 30° from the axis, the response level varies by no more than a couple of decibels. The response appears to rise slightly in the vicinity of 20°. This could, perhaps, be reduced by a more appropriate shaping of the sides of the reflector.

The variation of the response curve with the ratio $\frac{b}{a}$ has been studied briefly. It appears that a value of $\frac{b}{a} = 1$ is about right, for the 30° contour to equal the axial response. If $\frac{b}{a} < 1$, the reflector will only be partially corrected; if $\frac{b}{a} > 1$, it will be overly corrected. In the uncorrected reflector with triangular aperture, $\frac{b}{a} = 0$.

If $b/a = \infty$, that is, if $a = 0$, one would expect to obtain a response curve having a minimum value on the axis and rising to a maximum on either side. A reflector having these properties is illustrated in Figs. 12a and 12b. Again Fig. 12a is the axial aspect, whereas Fig. 12b is the 30° aspect. In the former, the effective area should be zero; in the latter, it has the value represented by the shaded portion. A reflector of this kind, in which b = 34″ and a = 0, was constructed and tested. The experimental results are shown in Fig. 13. The minimum is, perhaps, not as low in value as expected because of residual reflections from the support upon which the reflector was mounted. As expected, however, the curve passes from a minimum on the axis to maxima on either side.

The above examples serve to illustrate some of the results which can be realized with trihedral reflectors. We have seen that the response characteristic can be controlled by appropriate modifications of the geometrical configuration of the aperture.

Experiments were performed in order to determine the reduction in echo caused by errors in the internal angles at the corner.

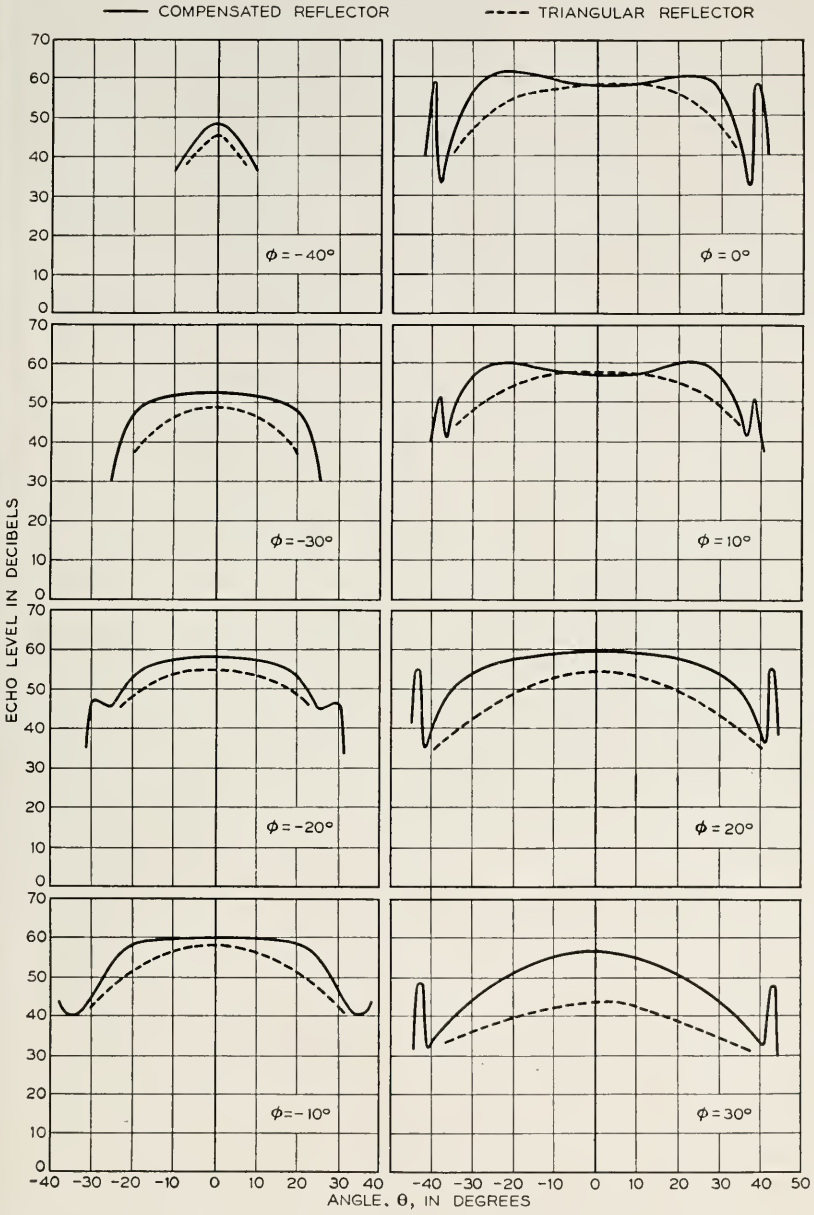


Fig. 11—Response of compensated reflector compared with that of triangular reflector.

In the first set of experiments only one of the internal angles was altered from its nominal value of 90°. Figure 14 shows the apparatus used in the

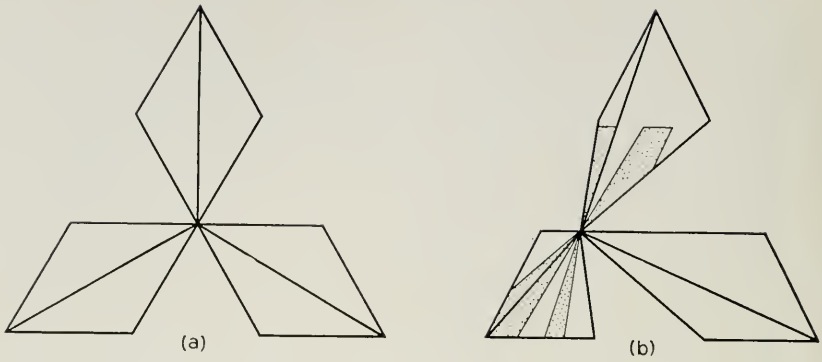


Fig. 12—Modified reflector having minimum response on axis.

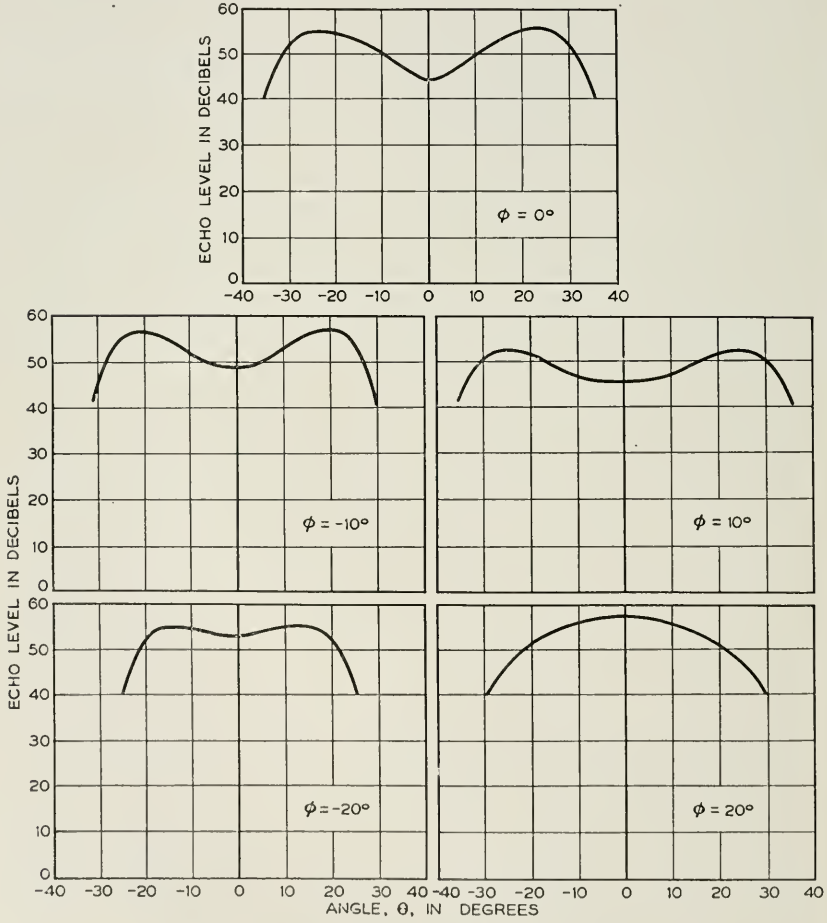


Fig. 13—Response patterns of reflector of Fig. 12.

experiment. The 24-inch reflector was constructed of silver-painted plywood and hinged along the intersection of the two upper surfaces so that the angle α could be varied at will. A series of response patterns were taken for various values of α. These are shown in the lower part of Fig. 14. It will be observed that one effect of changing α is to lower the echo level. This appears, however, to be accompanied by a somewhat flatter response curve. The radar used in this experiment had a wavelength of 1.25 centimeters.

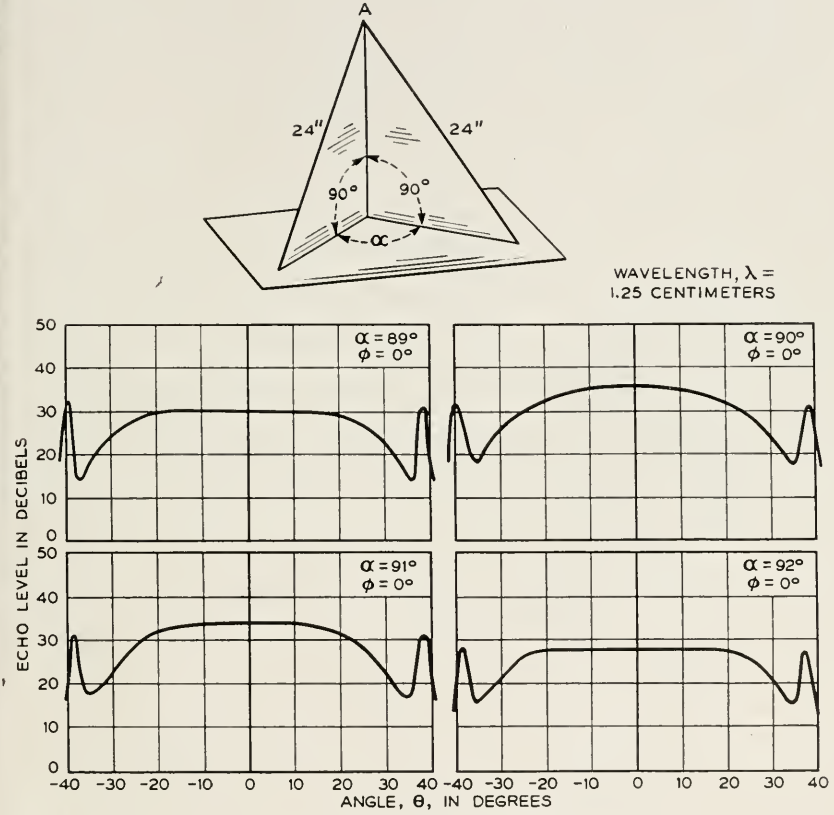


Fig. 14—Effect of an error in one of the corner angles of a trihedral upon its performance.

A second series of experiments was conducted in which all three internal angles were varied simultaneously. The curves shown in the upper half of Fig. 15 show the axial echo level for various sizes of reflector as a function of the internal angles α. It will be noted that for the 24-inch reflector an angular error of $\frac{1}{2}$ degree results in an echo-level reduction of two decibels, while the same angular error in the $9\frac{3}{8}$ inch reflector produces only a negligible reduction. The lower part of the figure shows some response patterns

taken with a 24-inch triangular reflector for several values of α. A wavelength of 1.25 centimeters was used in obtaining both sets of curves. Later, similar measurements were made at a wavelength of 3.2 centimeters. It was found that the loss of signal is a function of the linear error of the aperture in wavelengths rather than the angular error in degrees. Thus a given angular error in a $9\frac{3}{8}$-inch reflector at a wavelength of 1.25 centimeters will produce the same loss in signal as the same angular error in a 24-inch reflector operating at a wavelength of 3.2 centimeters.

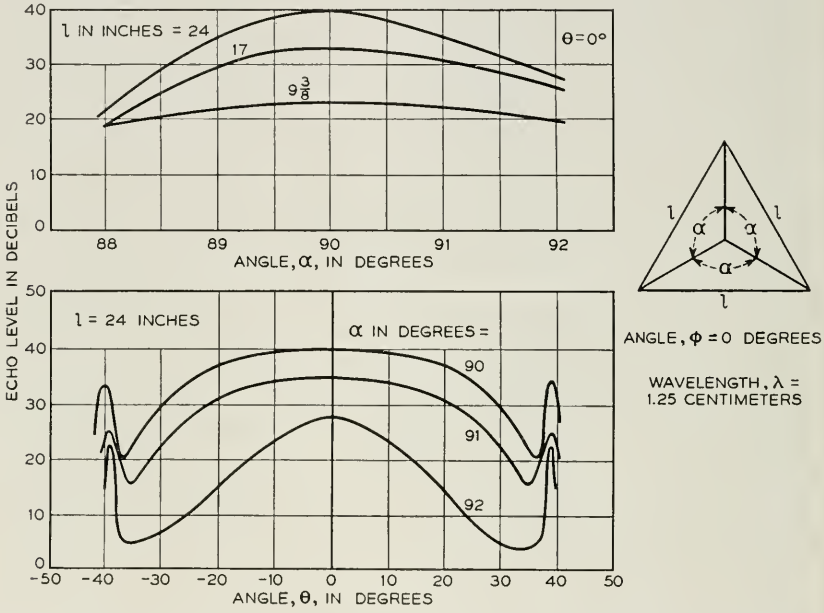


Fig. 15—Effect of an error in all three corner angles upon the performance of a trihedral.

Spheres and Cylinders

Formulas for the effective areas of spheres and cylinders which have dimensions large in comparison with a wavelength have been supplied by J. F. Carlson and S. A. Goudsmit of the Radiation Laboratory.[3] The effective area of a sphere of radius R is given by

$$A = \frac{R\lambda}{2} \tag{4}$$

where λ is the wavelength.

For a cylinder of radius R and height L, both large with respect to a wavelength, the effective area for rays perpendicular to the axis of the cylinder is

$$A = L\sqrt{\frac{R\lambda}{2}} \tag{5}$$

[3]Unpublished Report

It will be of interest to compare the sphere and the cylinder with corner reflectors and flat plates. The response pattern of a sphere is ideal in that it is uniform in all directions. Unfortunately its effective area is small in comparison with that of corner reflectors or flat plates having the same cross-sectional area. The cylinder has a symmetrical response pattern in the plane

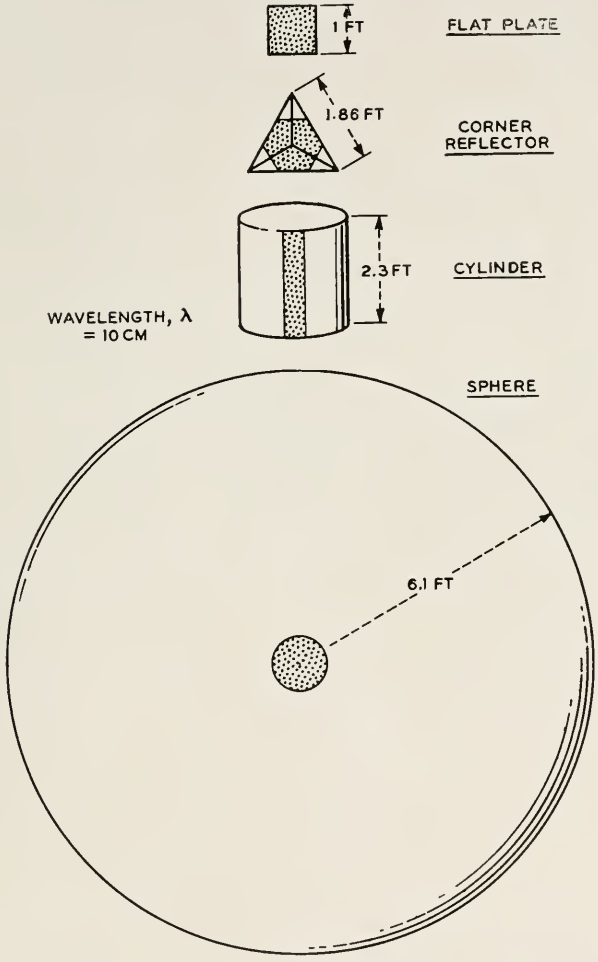


Fig. 16—A comparison of several representative targets having equal effective areas.

perpendicular to the axis but is very sharp in the plane of the axis. The effective area of a cylinder is intermediate between that of a corner reflector and a sphere. Figure 16 is a scale view of a flat plate, a corner reflector, a cylinder, and a sphere all having an effective area of one square foot at a wavelength of 10 centimeters. For shorter wavelengths the flat plate and

the corner reflector would remain the same size, whereas the cylinder and the sphere would have to be larger in order to maintain the same effective areas. At a wavelength of 1 centimeter the sphere would have to have a radius of about 60 feet.

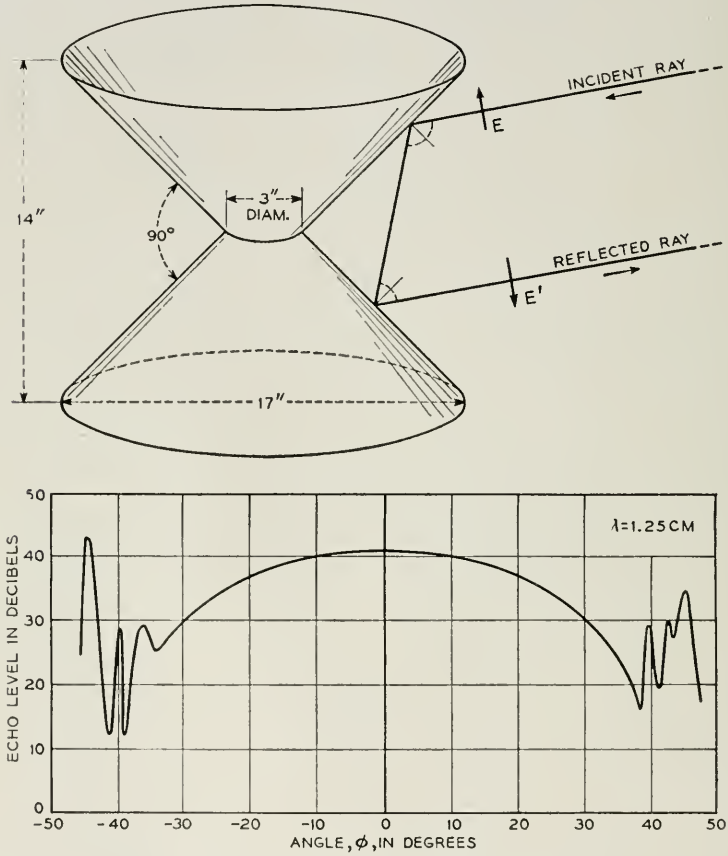


Fig. 17—Properties of the biconical corner reflector.

Biconical Corner Reflector

A reflector, combining the 360° horizontal response characteristic of the cylinder with a vertical response like that of the corner reflector, was evolved and is illustrated in Fig. 17. As shown here, the device consists of two conical surfaces placed in juxtaposition such that the generatrices of one cone intersect those of the other at right angles. The operation of the reflector is somewhat like that of a dihedral corner reflector in that a ray, upon striking one of the cones, is reflected to the other and then returns in the direction of

the source. The "Biconical" reflector may perhaps be likened to a cylinder which automatically orients itself so that the impinging rays are always perpendicular to the axis.

A biconical reflector was constructed of sheet metal having the dimensions indicated in Fig. 17. The vertical response pattern was measured and is plotted in the lower portion of the figure. Because of the circular symmetry of the reflector, the vertical response curve shown will be equally valid for all angles of azimuth. At $\phi = 0°$ the reflector exhibited an effective area of 0.16 square feet. The measurements were made at a wavelength of 1.25 centimeters.

In the above experiment the incident radiation was polarized in the plane of the axis of the cones. In another test with the polarization perpendicular to the axis the received echo was reduced by four decibels. This effect is not as yet entirely explained. It probably results from a depolarizing effect similar to that encountered in the dihedral corner reflector, complicated however by the curvature of the cones.

Only a limited amount of data is available for predicting the effective area of a biconical reflector over a wide range of sizes and wavelengths. The available data indicate that to a rough approximation and for a given polarization the effective area varies directly as the square root of the wavelength and as the three-halves power of the diameter of the cones, assuming that the height of the reflector is approximately equal to the diameter.

Tables of Phase Associated with a Semi-Infinite Unit Slope of Attenuation

By D. E. THOMAS

This paper presents tables of the phase associated with a semi-infinite unit slope of attenuation. The phase is given in degrees to .001 degree with an accuracy of ± .001 degree and in radians to .00001 radian with an accuracy of ± .000015 radian. The method of constructing the tables and a brief analysis of the errors are given. An appendix, which gives a detailed explanation with specific examples of the use of the tables in determining the phase associated with a given attenuation characteristic or the reactance associated with a given resistance characteristic by means of the straight line approximation method given in Bode's "Network Analysis and Feedback Amplifier Design," is included for the benefit of those who are not already acquainted with this method. The Appendix also presents an example of a non-minimum phase network[1] in which the minimum phase determined from the attenuation characteristic fails to predict the true phase of the network.

THE method described by Bode[2] for the determination of the phase associated with a given attenuation characteristic or the reactance associated with a given resistance characteristic has proved to be an extremely useful laboratory and design tool. In this method the attenuation (or real) characteristic, plotted versus the log of frequency, is approximated by a series of straight lines. The phase (or imaginary component) is then determined by summing up the individual contributions of each elementary straight line segment to the total phase (or imaginary component).

The most elementary straight line characteristic which can be used to construct a given straight line approximation is that in which the attenuation plotted against the log of frequency is constant on one side of a prescribed frequency, f_0, and has a constant slope thereafter. Such a characteristic has been called by Bode a "semi-infinite constant slope" characteristic.[3] A semi-infinite unit slope of attenuation or one in which the attenuation changes 6 *db* per octave, or 20 *db* per decade is shown in Fig. 1. The phase associated with this attenuation characteristic is plotted in Fig. 2.[4] The independent variable was chosen as f/f_0 for values of f less than f_0 and f_0/f for values of f greater than f_0 to keep it finite for all values of f and in order to show the phase plotted exactly as it is given in the tables to follow. The phase associated with a semi-infinite constant slope of

[1] For a complete discussion of *minimum phase* see Hendrik W. Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., New York, N. Y., 1945.

[2] Ibid: Chap. XV, page 344.

[3] Ibid: Chap. XIV, page 316.

[4] Ibid: Chap. XIV page 317.

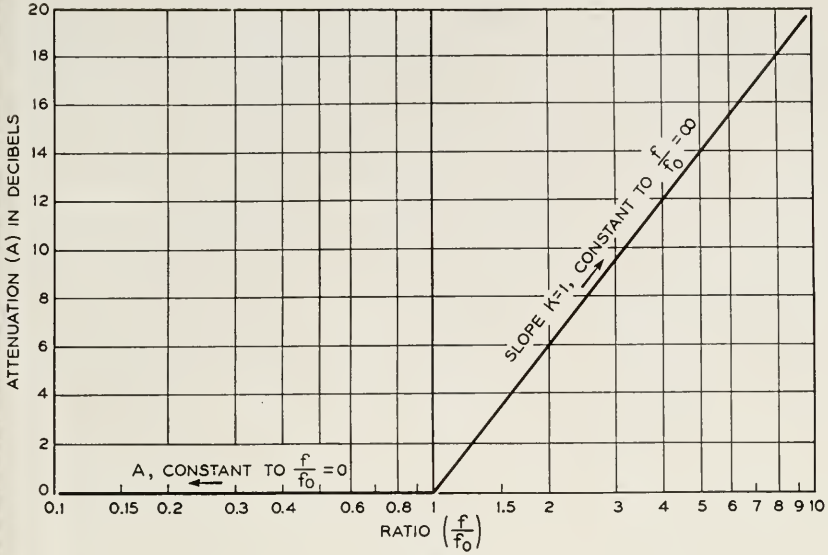


Fig. 1—Semi-infinite unit slope of attenuation.

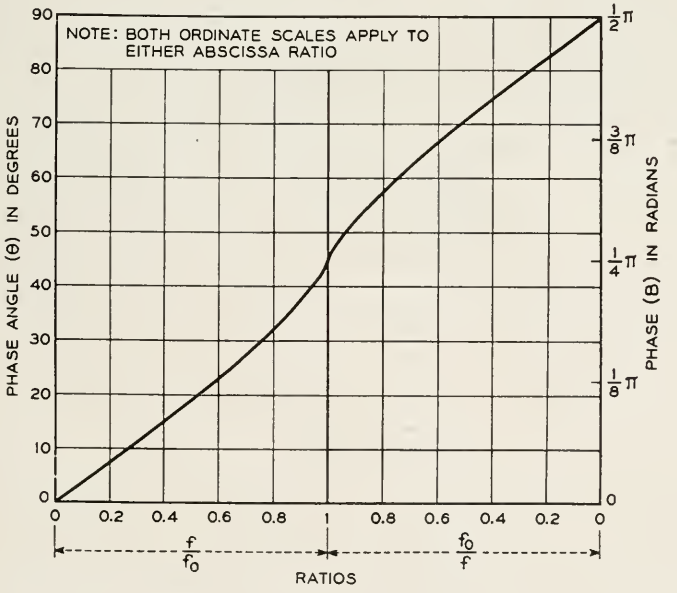


Fig. 2—Phase associated with semi-infinite unit slope of attenuation of Fig. 1.

attenuation of the same character as the semi-infinite unit slope of attenuation of Fig. 1 but of slope k, is k times the phase given in Fig. 2.

Bode points out,[5] however, that the building up of the complete imaginary characteristic from a single primitive curve, namely a semi-infinite real slope, suffers from the disadvantage that the phase contributions of the individual slopes may be rather large positive and negative quantities, even though the net phase shift is fairly small. In order to avoid this disadvantage, Bode recommends that the individual finite line segments which constitute the straight line approximation to the real characteristic be regarded as the elementary characteristics used in the summation of the total phase. He then gives a series of charts, plotted as a function of f/f_0, of the phase associated with a finite line segment having a 1 *db* change in attenuation and with a ratio of the geometric mean frequency (f_0) of the two terminal frequencies of the finite line segment to the lower terminal frequency as a parameter (ratio designated a).

However, problems have arisen where, even with the finite line segment phase charts, the phase contributions of the various elements were sufficiently large and nearly equal positive and negative quantities that difficulties in interpolation between the curves for the various values of a, given on the charts, resulted in a sufficient lack of precision that the quantity being sought was lost.

Because of the usefulness of the method in question, and with its application to a wider variety of problems, means of increasing its over-all precision and simplification of computation have constantly been sought. It had occurred to several engineers independently that a table of phase versus frequency for a semi-infinite unit slope of attenuation would prove extremely useful. The phase in radians at frequency f_c, associated with a semi-infinite unit slope of attenuation commencing at frequency f_0, is given by Bode as[6]

$$B(x_c) = \frac{2}{\pi}\left(x_c + \frac{x_c^3}{9} + \frac{x_c^5}{25} + \cdots\right) \tag{1}$$

where:

$$x_c = \frac{f_c}{f_0} = \frac{\omega_c}{\omega_0}, x_c < 1.$$

The computation time required to determine the phase at a given frequency by summation of the above series is such, that the work required to get the phase at a sufficient number of points and to a sufficient number of significant figures to prepare an adequate table proved to be sufficient to discourage this procedure.

[5] Ibid: Chap. XV page 338.
[6] Ibid: Chap. XV, page 343.

The derivative of (1) above, however, proves to be quite simple and easy to evaluate. It is given by Bode as:

$$\frac{dB}{dx_c} = \frac{1}{\pi x_c} \log \left| \frac{1 + x_c}{1 - x_c} \right| \tag{2}$$

$$= \frac{2}{\pi}\left(1 + \frac{x_c^2}{3} + \frac{x_c^4}{5} + \cdots\right), x_c < 1. \tag{2a}$$

It therefore seemed that since the phase had already been computed by the Mathematical Research Group of the Bell Telephone Laboratories, Inc., at a

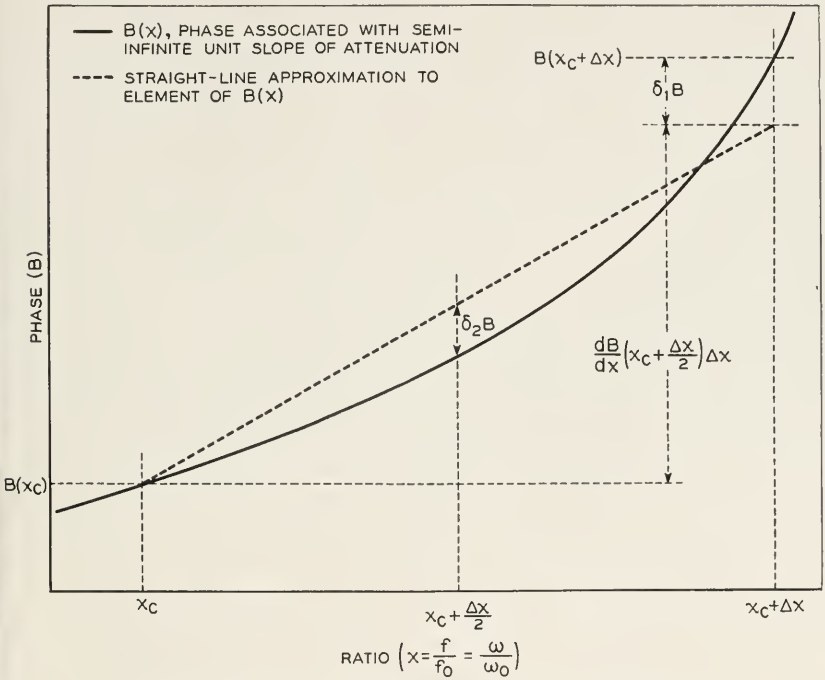


Fig. 3—Element of Fig. 2 for $f/f_0 < 1$ expanded qualitatively.

considerable number of points, using the infinite series expansion of (1) above the function in the regions between known values of phase could be constructed by assuming the intervening curve of phase as a function of $x = \frac{f}{f_0}$ to be a series of straight lines having the slope given by (2) above over intervals Δx of x made sufficiently small that the resultant straight line approximation would approach the true phase curve to the desired degree of accuracy for the table contemplated.

In order to evaluate the errors involved in such a procedure let us refer to Fig. 3 where a segment of the desired phase function to be constructed is qualitatively represented on a large scale. It is assumed that the phase at x_c, $B(x_c)$, is known and that it is desired to determine the error $\delta_1 B$ in phase computed for $x_c + \Delta x$ when it is assumed that the phase curve is a straight line from $B(x_c)$ at x_c, to $x_c + \Delta x$ having a slope, $\frac{dB}{dx}\left(x_c + \frac{\Delta x}{2}\right)$, the slope of the true phase curve at $x = x_c + \frac{\Delta x}{2}$.

Then:

$$\delta_1 B = B(x_c + \Delta x) - B(x_c) - \frac{dB}{dx}\left(x_c + \frac{\Delta x}{2}\right)\Delta x \tag{3}$$

where:

$$B(x_c) = \frac{2}{\pi}\left[x_c + \frac{x_c^3}{9} + \frac{x_c^5}{25} + \cdots\right]$$

$$B(x_c + \Delta x) = \frac{2}{\pi}\left[(x_c + \Delta x) + \tfrac{1}{9}(x_c^3 + 3x_c^2\,\Delta x + 3x_c\,\Delta x^2 + \Delta x^3)\right.$$
$$\left. + \tfrac{1}{25}(x_c^5 + 5x_c^4\,\Delta x + 10x_c^3\,\Delta x^2 + 10x_c^2\,\Delta x^3 + 5x_c\,\Delta x^4 + \Delta x^5) + \cdots\right]$$

$$B(x_c + \Delta x) - B(x_c) = \frac{2}{\pi}\left[\Delta x + \frac{x_c^2\,\Delta x}{3} + \frac{x_c\Delta x^2}{3}\right.$$
$$\left. + \frac{\Delta x^3}{9} + \frac{x_c^4\,\Delta x}{5} + \frac{2x_c^3\,\Delta x^2}{5} + \frac{2x_c^2\,\Delta x^3}{5} + \frac{x_c\Delta x^4}{5} + \frac{\Delta x^5}{25} + \cdots\right]$$
$$= \frac{2}{\pi}\left[\Delta x \sum_{n=1}^{n=\infty} \frac{x_c^{2n-2}}{2n-1} + \Delta x^2 \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{2n+1}\right.$$
$$\left. + \Delta x^3 \sum_{n=1}^{n=\infty} \frac{n(2n-1)x_c^{2n-2}}{3(2n+1)} + \cdots\right]$$

$$\frac{dB}{dx}\left(x_c + \frac{\Delta x}{2}\right)\Delta x = \frac{2}{\pi}\Delta x\left(1 + \frac{1}{3}\left[x_c^2 + 2x_c\left(\frac{\Delta x}{2}\right) + \left(\frac{\Delta x}{2}\right)^2\right]\right.$$
$$\left. + \frac{1}{5}\left[x_c^4 + 4x_c^3\left(\frac{\Delta x}{2}\right) + 6x_c^2\left(\frac{\Delta x}{2}\right)^2 + 4x_c\left(\frac{\Delta x}{2}\right)^3 + \left(\frac{\Delta x}{2}\right)^4\right] + \cdots\right)$$
$$= \frac{2}{\pi}\left[\Delta x + \frac{x_c^2\,\Delta x}{3} + \frac{x_c\Delta x^2}{3} + \frac{\Delta x^3}{12} + \frac{x_c^4\,\Delta x}{5} + \frac{2x_c^3\,\Delta x^2}{5}\right.$$
$$\left. + \frac{3x_c^2\,\Delta x^3}{10} + \frac{x_c\Delta x^4}{10} + \frac{\Delta x^5}{80} + \cdots\right]$$
$$= \frac{2}{\pi}\left[\Delta x \sum_{n=1}^{n=\infty} \frac{x_c^{2n-2}}{2n-1} + \Delta x^2 \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{2n+1}\right.$$
$$\left. + \Delta x^3 \sum_{n=1}^{n=\infty} \frac{n(2n-1)x_c^{2n-2}}{4(2n+1)} + \cdots\right].$$

Since Δx will be small compared to unity and since an error function is being computed it is permissible to take only the 1st term of the difference between the true phase and the computed phase, i.e. the Δx^3 term, and drop all higher order terms of Δx.
Then:

$$\delta_1 B \doteq \frac{2}{\pi}\left[\Delta x^3 \sum_{n=1}^{n=\infty} \frac{n(2n-1)x_c^{2n-2}}{3(2n+1)} - \Delta x^3 \sum_{n=1}^{n=\infty} \frac{n(2n-1)x_c^{2n-2}}{4(2n+1)}\right] = \frac{\Delta x^3}{6\pi} \sum_{n=1}^{n=\infty} \frac{n(2n-1)x_c^{2n-2}}{2n+1}. \tag{4}$$

The equation (4) above for $\delta_1 B$ gives only the error for a single increment Δx of $x = f/f_0$. If the phase is known at $x = x_a$ and $x = x_b$ and it is desired to determine the phase at points between $x = x_a$ and $x = x_b$ then since $\delta_1 B$ always has the same sign the errors due to successive increments of x will be cumulative and the total error at $x = x_b$ will be n times the average of the $\delta_1 B$ errors of each increment of Δx between x_a and x_b where n is the total number of equi-increments of x taken between x_a and x_b. However, since the individual $\delta_1 B$ errors decrease as the cube of Δx, the individual errors will decrease as the cube of the number of increments taken between the two frequencies at which the phase is known, whereas the cumulative $\delta_1 B$ error will increase only in proportion to the first power of n. Therefore, the net result will be a vanishing of the cumulative error inversely as the square of the number of frequency increments taken to approximate the curve in the interval in question. It therefore follows that the accuracy of the proposed method of building up the function, in so far as the phase at the terminals of the straight line segments is concerned, is limited only by the number of increments of frequency selected for the summation.

In order to determine the actual magnitude of errors to be expected $\delta_1 B$ was computed for $x_c = .4$ and $\Delta x = .02$ and found to be only .000015 degree. Since the total number of .02 intervals needed to be used between previously computed values of B is 5, the total cumulative error in this region for increments of this magnitude will not be greater than .0001 degree, which is entirely satisfactory, since the accuracy being sought is $\pm$ .0005 degree in B. For $x_c = .9$ and $\Delta x = .005$ the $\delta_1 B$ error proves to be only .00001 degree and since in this region the value of B has already been determined at .01 intervals by the more accurate series expansion technique referred to above, only two increments are necessary between known values of B and therefore the $\delta_1 B$ error is sufficiently small.

Having determined the order of magnitude of intervals necessary to keep $\delta_1 B$ errors small, let us examine the errors due to the departure of the straight line approximation from the true curve in the interval between x_c and $x_c + \Delta x$. Since $\delta_1 B$ will be very small it is anticipated that the maximum value

of $\delta_2 B$ (see Fig. 3) will occur in the vicinity of $x_c + \frac{\Delta x}{2}$. $\delta_2 B$ at this point may be determined as shown below.

$$\delta_2 B = B\left(x_c + \frac{\Delta x}{2}\right) - B(x_c) - \frac{dB}{dx}\left(x_c + \frac{\Delta x}{2}\right)\frac{\Delta x}{2} \tag{5}$$

where:

$$B\left(x_c + \frac{\Delta x}{2}\right) - B(x_c) = \frac{2}{\pi}\left[\Delta x \sum_{n=1}^{n=\infty} \frac{x_c^{2n-2}}{2(2n-1)} + \Delta x^2 \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{4(2n+1)} + \cdots\right]$$

$$\frac{dB}{dx}\left(x_c + \frac{\Delta x}{2}\right)\frac{\Delta x}{2} = \frac{2}{\pi}\left[\Delta x \sum_{n=1}^{n=\infty} \frac{x_c^{2n-2}}{2(2n-1)} + \Delta x^2 \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{2(2n+1)} + \cdots\right].$$

Again retaining only the first term of the error function and dropping all higher order terms of Δx

$$\begin{aligned}\delta_2 B &\doteq \frac{2}{\pi}\left[\Delta x^2 \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{4(2n+1)} - \Delta x^2 \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{2(2n+1)}\right] \\ &= -\frac{\Delta x^2}{2\pi} \sum_{n=1}^{n=\infty} \frac{n x_c^{2n-1}}{2n+1}.\end{aligned} \tag{6}$$

$\delta_2 B$ proves to be negative and considerably larger than $\delta_1 B$ for the same magnitude of interval. Therefore the computed B will always exceed the true phase in the interval x_c to $x_c + \Delta x$ except above a value of x very near to $x_c + \Delta x$ where the straight line approximation crosses the true phase curve. When $x_c = .35$ and $\Delta x = .02$, $\delta_2 B$ is found to be $-.0005$ degree from (6) above, and for $x_c = .91$ and $\Delta x = .005$, $\delta_2 B$ is also found to be $-.0005$ degree. The $\delta_2 B$ errors are therefore found to be much more important than the $\delta_1 B$ errors. $\delta_2 B$ errors are not accumulative, however, and therefore increments of Δx of the above order of magnitude prove to be sufficiently small to give the accuracy being sought, namely $\pm .0005$ degree in B.

An evaluation of the $\delta_1 B$ and $\delta_2 B$ errors for values of x_c greater than .9 is difficult due to the slowness of convergence of the series giving these errors. For values of x_c between .9 and unity, however, the frequency of known values of B determined from (1) above and available as check points is sufficient to check the adequacy of intervals insofar as $\delta_1 B$ errors are concerned. Furthermore an analysis similar to that given above for the determination of the $\delta_1 B$ and $\delta_2 B$ errors shows that an interpolation of the slopes computed for construction of the tables in question, to give the intervening slopes necessary to cut the increments of Δx in half will give check points at $x_c + \frac{\Delta x}{2}$ frequencies, with a $\delta_1 B$ error ($x_c + \frac{\Delta x}{2}$ is then the termination of a straight

line segment since the Δx interval has been halved) of comparable order of magnitude to the $\delta_1 B$ error for the original interval selected and therefore small in comparison to the $\delta_2 B$ error for the original Δx interval. This technique was therefore used in checking the adequacy of the intervals in so far as $\delta_2 B$ errors are concerned in the region $x_c = .9$ to $x_c = 1.0$.

Using the procedure outlined above the phase associated with the semi-infinite unit slope of attenuation of Fig. 1 was computed for values of f less than f_0 and is given as a function of f/f_0 in Table I in degrees and in Table III in radians. For values of f greater than f_0 the phase was computed as a function of f_0/f utilizing the odd symmetry behavior of the phase characteristic of Fig. 2 on opposite sides of $f/f_0 = 1$, and this phase is tabulated in Table II in degrees and in Table IV in radians. For the other type of semi-infinite unit slope of attenuation in which the attenuation slope is constant and equal to unity at all frequencies below f_0 and the attenuation is constant for all frequencies above f_0 (with the constant slope of attenuation intersecting the f_0 axis at the same point as the constant attenuation line) the same tables can be used by reading the values of phase for $f/f_0 < 1$ from the f_0/f tables and the values of phase for $f_0/f < 1$ from the f/f_0 tables.

The intervals over which the straight line approximation to the true phase was assumed are given below:

Δx			x_c	
.02	from	.00	to	.40
.01	"	.40	"	.70
.005	"	.70	"	.92
.002	"	.92	"	.98
.001	"	.98	"	.996
.0005	"	.996	"	.998
.0002	"	.998	"	.999
.0001	"	.999	"	.9998
.00005	"	.9998	"	1.0000

The points at which the cumulative sum of the straight line increments of phase was corrected to the phase as determined from (1) above are listed below:

Every	.1	from	.00	to	.40
"	.05	"	.40	"	.80
"	.02	"	.80	"	.90
"	.01	"	.90	"	.99

and at .996, .998, .999, and 1.000

A study of the errors based on the error analysis discussed above indicates that the computed values of B in degrees are accurate to $\pm$ .0005 degree and since there is an additional possibility of $\pm$ .0005 degree error in dropping all figures beyond the third decimal place, the over-all reliability of the degree tables is $\pm$ .001 degree. Similarly the computed values of B in radians are accurate to $\pm$ .00001 radian and since there is an additional possibility of $\pm$ .000005 radian error in dropping all figures beyond the fifth decimal

place, the over-all reliability of the radian tables is ± .000015 radian. Since the function tabulated was constructed by a series of straight line approximations to the true phase, interpolation to get the phase for values of f/f_0 or f_0/f between those given in the tables in problems where this is necessary, will result in the same accuracy as that given for the tabulated values.

Murlan S. Corrington[7] of Radio Corporation of America has computed the phase in radians for the semi-infinite unit slope of attenuation of Fig. 1 for approximately 100 values of f/f_0 using equations 15-9 and 15-11 of Bode's "Network Analysis and Feedback Amplifier Design" and has given a table of these values to five decimal places. Where the values of Table III differ from Corrington's values, his value is given as a superscript. Since his approach is the more exact one, it is assumed that where a difference exists, his value is correct. The differences have a maximum value of one figure in the fifth decimal place which is consistent with the accuracy of ± .000015 radian given for Table III. However, linear interpolation of Corrington's values to get the function to three figures in f/f_0, which precision in f/f_0 is really needed to utilize five figure accuracy in B, will result in errors considerably larger than those of Table III for the higher values of f/f_0.

Acknowledgment

The writer wishes to thank Miss J. D. Goeltz who carried out the calculations of the basic Tables and of the illustrative examples of this paper.

APPENDIX

Use of Tables I to IV in Determining Phase from Attenuation or Reactance from Resistance

The first step in determining the phase associated with a given attenuation characteristic using the tables described in the basic paper is to plot the attenuation as a function of log frequency to a suitable scale. Such an attenuation characteristic is illustrated in Fig. 4a. The attenuation characteristic is then approximated by a series of straight lines such as are shown in dotted form. The number of straight lines used will depend upon the accuracy desired in the resultant phase. As a rule, an approximation to the attenuation which does not depart by more than ± .5 *db* will give a resultant phase which does not depart by more than ± 3° from the true phase.

If we now examine the straight line attenuation approximation of Fig. 4a,

[7] Murlan S. Corrington, "Table of the Integral $\frac{2}{\pi}\int_0^\infty \frac{\tanh^{-1} t}{t}\, dt$" *R.C.A. Review* September, 1946, page 432.

we see that it can be constructed by adding a number of semi-infinite constant slopes of attenuation as shown in Fig. 4b. The first of these will be a semi-infinite slope of magnitude k_1 commencing at the first critical frequency

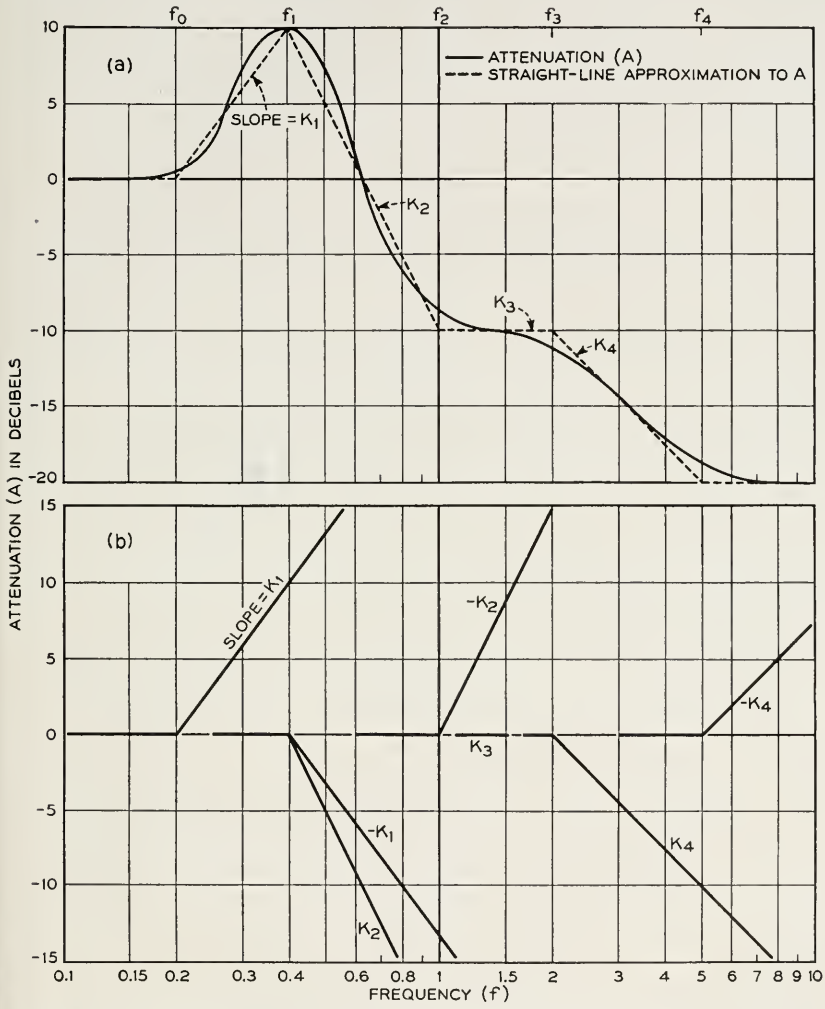


Fig. 4—(a) Straight line approximation to attenuation characteristic. (b) Individual semi-infinite constant slopes of attenuation which add to produce the straight line approximation of Fig. 4(a).

f_0. The second will be a semi-infinite slope of magnitude $-k_1$ commencing at the critical frequency f_1 which must be added to correct for the fact that the first straight line of slope $+k_1$ does not extend to infinity, but terminates at the critical frequency f_1, where the straight line approximation assumes a

new slope. In order to achieve this new slope a semi-infinite slope of magnitude k_2, commencing at frequency f_1, must be added. This process is continued up the frequency scale until the entire straight line approximation is constructed.

The total phase $\theta(f)$ at a particular frequency f is then given by the sum of the phase at frequency f associated with each of the semi-infinite constant slopes of attenuation which together make up the straight line approxmation.

Thus:

$$\theta(f) = k_1\theta_0 - k_1\theta_1 + k_2\theta_1 - k_2\theta_2 + k_3\theta_2 - k_3\theta_3 + k_4\theta_3 - k_4\theta_4$$

or for the general straight line approximation having slopes

$$k_1 , k_2 , \cdots k_n$$

$$\theta(f) = k_1 (\theta_0 - \theta_1) + k_2 (\theta_1 - \theta_2) + \cdots k_n (\theta_{n-1} - \theta_n)$$

where:

θ_n is the phase at frequency f associated with the semi-infinite unit slope of attenuation commencing at frequency f_n and extending to $f = \infty$ and is read from Tables I or III for $f < f_n$ and Tables II or IV for $f > f_n$,

and

k_n is the slope of the straight line approximation between f_{n-1} and f_n given by:

$$k_n = \frac{A_n - A_{n-1}}{20 \log \frac{f_n}{f_{n-1}}}$$

where:

A_n is the attenuation at frequency f_n on the straight line approximation.

Note that in Fig. 4a the attenuation is constant from zero frequency to the first critical frequency f_0. In many problems, there is a constant slope below frequency f_1 to frequency zero. In that event, the initial critical frequency, f_0, will be zero, and θ_0 will be 90°. ($f_0/f = 0$ at all finite frequencies.) When this occurs, k_1 must be determined by choosing a finite frequency f_0' and taking the ratio of attenuation change between f_0' and f_1 to 20 log of the ratio of f_1 to f_0'. Similarly, the attenuation is constant in the illustration from the top critical frequency f_4 to infinity, whereas in many problems the attenuation will have a constant slope extending from the top critical frequency to infinity. In these cases, the top critical frequency will be infinity and the final angle θ_n will, of course, be zero. Here again the final slope k_n must be determined over a finite portion of this infinite slope.

It will also be noted that in the illustration given the characteristic is approximated, commencing at zero frequency, by a series of semi-infinite slopes, each of which is a constant times the characteristic of Fig. 1 of the basic paper, for which Tables I to IV were computed. The characteristic could have been approximated just as well with a series of semi-infinite constant slopes, commencing at $f = \infty$ and going down in frequency, each having a flat attenuation above a critical frequency f_n and constant slope at frequencies below. In summing the phase for such an approximation Tables I to IV may be used by reading the angles for f/f_n from the f_0/f tables and vice versa as indicated in the basic paper.

As an illustration of the above procedure, consider the determination of the phase associated with the characteristic given by 20 log | Z | shown in Fig. 5. The characteristic is first approximated by a series of straight lines as shown in dotted form. The critical frequencies and values of $A = 20$ log | Z | at these critical frequencies are then read from the straight line approximation[8] and the slopes of the various straight line segments determined as illustrated in Table V.

Having determined the slopes of the various segments of the straight line approximation, the phase at any desired frequency is summed as illustrated in Table VI where the phase for $f = 1.5$ is summed.

The mesh computed value of θ for the network in question is plotted in Fig. 6 and it will be noted that the phase summation of Table VI checks the true value to within the accuracy to which the phase can be read from the curve. The identical procedure is followed in determining the phase at any other frequency. As an illustration of the accuracy of the method, the phase was determined at a considerable number of frequencies and the results shown as individual points in Fig. 6. The straight line approximation to 20 log | Z | of Fig. 5 was of the order of $\pm$ .25 *db* and, in accordance with the estimated accuracy of the method given above, the maximum departure of the phase summation from the true phase is approximately $\pm$ 1.5°.

A much simpler approximation than that of Fig. 5 may be used without a great loss in accuracy. For instance, a five-line approximation determined by the critical frequencies of Table VII will match 20 log | Z | to within approximately $\pm$ .5 *db* and therefore should give a phase summation within $\pm$ 3° of the true phase. The phase was actually summed at 12 frequencies chosen at random for this five-line approximation and the maximum departure of the summed phase from the true phase was 3.2°. With experience in use of the method, simpler approximations can be used and the phase determined more accurately than the limits of accuracy of the summation at individual frequencies by plotting the individual summations

[8] The original plot was expanded and had much greater scale detail than can be shown with clarity on a single page plate.

TABLE I—DEGREES PHASE (±.001°) FOR SEMI-INFINITE ATTENUATION SLOPE $k = 1 f < f_0$

f/f_0	0	1	2	3	4	5	6	7	8	9
.00	.000	.036	.073	.109	.146	.182	.219	.255	.292	.328
.01	.365	.401	.438	.474	.511	.547	.584	.620	.657	.693
.02	.730	.766	.803	.839	.875	.912	.948	.985	1.021	1.058
.03	1.094	1.131	1.167	1.204	1.240	1.277	1.313	1.350	1.386	1.423
.04	1.459	1.496	1.532	1.569	1.605	1.642	1.678	1.715	1.751	1.788
.05	1.824	1.861	1.897	1.934	1.970	2.007	2.043	2.080	2.116	2.153
.06	2.189	2.226	2.262	2.299	2.335	2.372	2.409	2.445	2.482	2.518
.07	2.555	2.591	2.628	2.664	2.701	2.737	2.774	2.810	2.847	2.884
.08	2.920	2.957	2.993	3.030	3.066	3.103	3.140	3.176	3.213	3.249
.09	3.286	3.322	3.359	3.396	3.432	3.469	3.505	3.542	3.578	3.615
.10	3.652	3.688	3.725	3.762	3.798	3.835	3.871	3.908	3.945	3.981
.11	4.018	4.054	4.091	4.128	4.164	4.201	4.238	4.274	4.311	4.347
.12	4.384	4.421	4.457	4.494	4.531	4.568	4.604	4.641	4.678	4.714
.13	4.751	4.788	4.824	4.861	4.898	4.934	4.971	5.008	5.044	5.081
.14	5.118	5.155	5.191	5.228	5.265	5.302	5.338	5.375	5.412	5.449
.15	5.485	5.522	5.559	5.596	5.632	5.669	5.706	5.743	5.779	5.816
.16	5.853	5.890	5.927	5.963	6.000	6.037	6.074	6.111	6.148	6.184
.17	6.221	6.258	6.295	6.332	6.369	6.405	6.442	6.479	6.516	6.553
.18	6.590	6.626	6.663	6.700	6.737	6.774	6.811	6.848	6.885	6.922
.19	6.959	6.996	7.033	7.070	7.106	7.143	7.180	7.217	7.254	7.291
.20	7.328	7.365	7.402	7.439	7.476	7.513	7.550	7.587	7.624	7.661
.21	7.698	7.735	7.772	7.809	7.846	7.883	7.920	7.957	7.994	8.032
.22	8.069	8.106	8.143	8.180	8.217	8.254	8.291	8.329	8.366	8.403
.23	8.440	8.477	8.514	8.551	8.589	8.626	8.663	8.700	8.737	8.774
.24	8.811	8.849	8.886	8.923	8.960	8.998	9.035	9.072	9.109	9.147
.25	9.184	9.221	9.259	9.296	9.333	9.370	9.408	9.445	9.482	9.519
.26	9.557	9.594	9.631	9.669	9.706	9.744	9.781	9.818	9.856	9.893
.27	9.931	9.968	10.006	10.043	10.080	10.118	10.155	10.193	10.230	10.267
.28	10.305	10.342	10.380	10.417	10.455	10.492	10.530	10.568	10.605	10.643
.29	10.680	10.718	10.755	10.793	10.830	10.868	10.906	10.943	10.981	11.018
.30	11.056	11.094	11.131	11.169	11.207	11.244	11.282	11.320	11.358	11.395
.31	11.433	11.471	11.508	11.546	11.584	11.622	11.659	11.697	11.735	11.772
.32	11.810	11.848	11.886	11.924	11.962	12.000	12.037	12.075	12.113	12.151
.33	12.189	12.227	12.265	12.303	12.341	12.379	12.416	12.454	12.492	12.530
.34	12.568	12.606	12.644	12.682	12.720	12.758	12.797	12.835	12.873	12.911
.35	12.949	12.987	13.025	13.063	13.101	13.139	13.177	13.215	13.254	13.292
.36	13.330	13.368	13.406	13.445	13.483	13.521	13.559	13.598	13.636	13.674
.37	13.713	13.751	13.789	13.827	13.866	13.904	13.942	13.981	14.019	14.057
.38	14.096	14.134	14.173	14.211	14.250	14.288	14.327	14.365	14.404	14.442
.39	14.481	14.519	14.558	14.596	14.635	14.673	14.712	14.750	14.789	14.827
.40	14.866	14.905	14.943	14.982	15.021	15.059	15.098	15.137	15.175	15.214
.41	15.253	15.292	15.330	15.369	15.408	15.447	15.486	15.525	15.563	15.602
.42	15.641	15.680	15.719	15.758	15.797	15.836	15.875	15.914	15.953	15.991
.43	16.030	16.070	16.109	16.148	16.187	16.226	16.265	16.304	16.343	16.382
.44	16.421	16.460	16.500	16.539	16.578	16.617	16.657	16.696	16.735	16.774
.45	16.813	16.853	16.892	16.931	16.971	17.010	17.050	17.089	17.128	17.168
.46	17.207	17.247	17.286	17.326	17.365	17.405	17.444	17.484	17.523	17.563
.47	17.602	17.642	17.681	17.721	17.761	17.800	17.840	17.880	17.919	17.959
.48	17.999	18.039	18.078	18.118	18.158	18.198	18.238	18.277	18.317	18.357
.49	18.397	18.437	18.477	18.517	18.557	18.597	18.637	18.677	18.717	18.757
.50	18.797	18.837	18.877	18.917	18.958	18.998	19.038	19.078	19.118	19.158
.51	19.198	19.239	19.279	19.320	19.360	19.400	19.441	19.481	19.521	19.562
.52	19.602	19.642	19.683	19.723	19.764	19.804	19.845	19.885	19.926	19.967
.53	20.007	20.048	20.088	20.129	20.170	20.211	20.251	20.292	20.333	20.373
.54	20.414	20.455	20.496	20.537	20.578	20.619	20.660	20.701	20.741	20.782

TABLE I—*Continued*

f/f_0	0	1	2	3	4	5	6	7	8	9
.55	20.823	20.864	20.906	20.947	20.988	21.029	21.070	21.111	21.152	21.193
.56	21.234	21.276	21.317	21.358	21.400	21.441	21.482	21.524	21.565	21.606
.57	21.648	21.689	21.731	21.772	21.814	21.855	21.897	21.939	21.980	22.022
.58	22.063	22.105	22.147	22.189	22.230	22.272	22.314	22.356	22.397	22.439
.59	22.481	22.523	22.565	22.607	22.649	22.691	22.733	22.775	22.817	22.859
.60	22.901	22.943	22.986	23.028	23.070	23.112	23.155	23.197	23.239	23.281
.61	23.324	23.366	23.409	23.451	23.494	23.536	23.579	23.621	23.664	23.706
.62	23.749	23.792	23.834	23.877	23.920	23.963	24.006	24.048	24.091	24.134
.63	24.177	24.220	24.263	24.306	24.349	24.392	24.435	24.478	24.521	24.564
.64	24.607	24.651	24.694	24.738	24.781	24.824	24.868	24.911	24.954	24.998
.65	25.041	25.085	25.128	25.172	25.216	25.259	25.303	25.347	25.390	25.434
.66	25.478	25.522	25.566	25.610	25.654	25.698	25.742	25.786	25.830	25.873
.67	25.917	25.962	26.006	26.050	26.095	26.139	26.183	26.228	26.272	26.316
.68	26.361	26.405	26.450	26.494	26.539	26.584	26.628	26.673	26.718	26.762
.69	26.807	26.852	26.897	26.942	26.987	27.032	27.077	27.122	27.167	27.212
.70	27.257	27.302	27.348	27.393	27.438	27.484	27.529	27.574	27.620	27.665
.71	27.711	27.757	27.802	27.848	27.894	27.939	27.985	28.031	28.077	28.123
.72	28.169	28.215	28.261	28.307	28.353	28.399	28.445	28.492	28.538	28.584
.73	28.631	28.677	28.724	28.770	28.817	28.863	28.910	28.957	29.003	29.050
.74	29.097	29.144	29.191	29.238	29.285	29.332	29.379	29.426	29.473	29.521
.75	29.568	29.615	29.663	29.710	29.757	29.805	29.853	29.900	29.948	29.996
.76	30.043	30.091	30.139	30.187	30.235	30.283	30.331	30.379	30.428	30.476
.77	30.524	30.572	30.621	30.669	30.718	30.766	30.815	30.864	30.913	30.961
.78	31.010	31.059	31.108	31.157	31.206	31.255	31.305	31.354	31.403	31.453
.79	31.502	31.551	31.601	31.651	31.700	31.750	31.800	31.850	31.900	31.950
.80	32.000	32.050	32.100	32.150	32.201	32.251	32.301	32.352	32.403	32.453
.81	32.504	32.555	32.606	32.657	32.707	32.758	32.810	32.861	32.912	32.963
.82	33.015	33.066	33.118	33.170	33.221	33.273	33.325	33.377	33.429	33.481
.83	33.533	33.586	33.638	33.690	33.743	33.795	33.848	33.901	33.954	34.006
.84	34.059	34.113	34.166	34.219	34.272	34.325	34.379	34.433	34.486	34.540
.85	34.594	34.648	34.702	34.756	34.810	34.865	34.919	34.974	35.028	35.083
.86	35.138	35.193	35.248	35.303	35.358	35.413	35.469	35.524	35.580	35.636
.87	35.691	35.747	35.804	35.860	35.916	35.972	36.029	36.086	36.142	36.199
.88	36.256	36.313	36.370	36.428	36.485	36.542	36.600	36.658	36.716	36.774
.89	36.832	36.891	36.949	37.008	37.067	37.125	37.184	37.244	37.303	37.362
.90	37.422	37.482	37.542	37.602	37.662	37.722	37.783	37.844	37.904	37.965
.91	38.026	38.088	38.149	38.211	38.273	38.334	38.397	38.459	38.522	38.584
.92	38.647	38.710	38.773	38.837	38.901	38.965	39.029	39.093	39.157	39.222
.93	39.287	39.352	39.418	39.483	39.549	39.615	39.681	39.748	39.815	39.882
.94	39.949	40.017	40.085	40.153	40.221	40.290	40.359	40.428	40.497	40.567
.95	40.638	40.708	40.779	40.850	40.921	40.993	41.066	41.138	41.211	41.285
.96	41.358	41.432	41.507	41.582	41.657	41.733	41.809	41.887	41.964	42.042
.97	42.120	42.199	42.278	42.359	42.439	42.521	42.603	42.686	42.769	42.854
.98	42.938	43.024	43.111	43.199	43.288	43.378	43.469	43.561	43.655	43.750
.99	43.846	43.945	44.045	44.148	44.253	44.361	44.473	(refer to table below)		

.9960	44.473	.9984	44.763	.9992	44.871
.9965	44.530	.9985	44.776	.9993	44.886
.9970	44.589	.9986	44.789	.9994	44.900
.9975	44.649	.9987	44.802	.9995	44.915
.9980	44.711	.9988	44.816	.9996	44.931
.9981	44.724	.9989	44.829	.9997	44.946
.9982	44.737	.9990	44.843	.9998	44.963
.9983	44.750	.9991	44.857	.9999	44.980
				1.0000	45.000

TABLE II—DEGREES PHASE (±.001°) FOR SEMI-INFINITE ATTENUATION SLOPE $k = 1$
$f > f_0$

f_0/f	0	1	2	3	4	5	6	7	8	9
.00	90.000	89.964	89.927	89.891	89.854	89.818	89.781	89.745	89.708	89.672
.01	89.635	89.599	89.562	89.526	89.489	89.453	89.416	89.380	89.343	89.307
.02	89.270	89.234	89.197	89.161	89.125	89.088	89.052	89.015	88.979	88.942
.03	88.906	88.869	88.833	88.796	88.760	88.723	88.687	88.650	88.614	88.577
.04	88.541	88.504	88.468	88.431	88.395	88.358	88.322	88.285	88.249	88.212
.05	88.176	88.139	88.103	88.066	88.030	87.993	87.957	87.920	87.884	87.847
.06	87.811	87.774	87.738	87.701	87.665	87.628	87.591	87.555	87.518	87.482
.07	87.445	87.409	87.372	87.336	87.299	87.263	87.226	87.190	87.153	87.116
.08	87.080	87.043	87.007	86.970	86.934	86.897	86.860	86.824	86.787	86.751
.09	86.714	86.678	86.641	86.604	86.568	86.531	86.495	86.458	86.422	86.385
.10	86.348	86.312	86.275	86.238	86.202	86.165	86.129	86.092	86.055	86.019
.11	85.982	85.946	85.909	85.872	85.836	85.799	85.762	85.726	85.689	85.653
.12	85.616	85.579	85.543	85.506	85.469	85.432	85.396	85.359	85.322	85.286
.13	85.249	85.212	85.176	85.139	85.102	85.066	85.029	84.992	84.956	84.919
.14	84.882	84.845	84.809	84.772	84.735	84.698	84.662	84.625	84.588	84.551
.15	84.515	84.478	84.441	84.404	84.368	84.331	84.294	84.257	84.221	84.184
.16	84.147	84.110	84.073	84.037	84.000	83.963	83.926	83.889	83.852	83.816
.17	83.779	83.742	83.705	83.668	83.631	83.595	83.558	83.521	83.484	83.447
.18	83.410	83.374	83.337	83.300	83.263	83.226	83.189	83.152	83.115	83.078
.19	83.041	83.004	82.967	82.930	82.894	82.857	82.820	82.783	82.746	82.709
.20	82.672	82.635	82.598	82.561	82.524	82.487	82.450	82.413	82.376	82.339
.21	82.302	82.265	82.228	82.191	82.154	82.117	82.080	82.043	82.006	81.968
.22	81.931	81.894	81.857	81.820	81.783	81.746	81.709	81.671	81.634	81.597
.23	81.560	81.523	81.486	81.449	81.411	81.374	81.337	81.300	81.263	81.226
.24	81.189	81.151	81.114	81.077	81.040	81.002	80.965	80.928	80.891	80.853
.25	80.816	80.779	80.741	80.704	80.667	80.630	80.592	80.555	80.518	80.481
.26	80.443	80.406	80.369	80.331	80.294	80.256	80.219	80.182	80.144	80.107
.27	80.069	80.032	79.994	79.957	79.920	79.882	79.845	79.807	79.770	79.733
.28	79.695	79.658	79.620	79.583	79.545	79.508	79.470	79.432	79.395	79.357
.29	79.320	79.282	79.245	79.207	79.170	79.132	79.094	79.057	79.019	78.982
.30	78.944	78.906	78.869	78.831	78.793	78.756	78.718	78.680	78.642	78.605
.31	78.567	78.529	78.492	78.454	78.416	78.378	78.341	78.303	78.265	78.228
.32	78.190	78.152	78.114	78.076	78.038	78.000	77.963	77.925	77.887	77.849
.33	77.811	77.773	77.735	77.697	77.659	77.621	77.584	77.546	77.508	77.470
.34	77.432	77.394	77.356	77.318	77.280	77.242	77.203	77.165	77.127	77.089
.35	77.051	77.013	76.975	76.937	76.899	76.861	76.823	76.785	76.746	76.708
.36	76.670	76.632	76.594	76.555	76.517	76.479	76.441	76.402	76.364	76.326
.37	76.287	76.249	76.211	76.173	76.134	76.096	76.058	76.019	75.981	75.943
.38	75.904	75.866	75.827	75.789	75.750	75.712	75.673	75.635	75.596	75.558
.39	75.519	75.481	75.442	75.404	75.365	75.327	75.288	75.250	75.211	75.173
.40	75.134	75.095	75.057	75.018	74.979	74.941	74.902	74.863	74.825	74.786
.41	74.747	74.708	74.670	74.631	74.592	74.553	74.514	74.475	74.437	74.398
.42	74.359	74.320	74.281	74.242	74.203	74.164	74.125	74.086	74.047	74.009
.43	73.970	73.930	73.891	73.852	73.813	73.774	73.735	73.696	73.657	73.618
.44	73.579	73.540	73.500	73.461	73.422	73.383	73.343	73.304	73.265	73.226
.45	73.187	73.147	73.108	73.069	73.029	72.990	72.950	72.911	72.872	72.832
.46	72.793	72.753	72.714	72.674	72.635	72.595	72.556	72.516	72.477	72.437
.47	72.398	72.358	72.319	72.279	72.239	72.200	72.160	72.120	72.081	72.041
.48	72.001	71.961	71.922	71.882	71.842	71.802	71.762	71.723	71.683	71.643
.49	71.603	71.563	71.523	71.483	71.443	71.403	71.363	71.323	71.283	71.243
.50	71.203	71.163	71.123	71.083	71.042	71.002	70.962	70.922	70.882	70.842
.51	70.802	70.761	70.721	70.680	70.640	70.600	70.559	70.519	70.479	70.438
.52	70.398	70.358	70.317	70.277	70.236	70.196	70.155	70.115	70.074	70.033
.53	69.993	69.952	69.912	69.871	69.830	69.789	69.749	69.708	69.667	69.627
.54	69.586	69.545	69.504	69.463	69.422	69.381	69.340	69.299	69.259	69.218

TABLE II—*Continued*

f_0/f	0	1	2	3	4	5	6	7	8	9
.55	69.177	69.136	69.094	69.053	69.012	68.971	68.930	68.889	68.848	68.807
.56	68.766	68.724	68.683	68.642	68.600	68.559	68.518	68.476	68.435	68.394
.57	68.352	68.311	68.269	68.228	68.186	68.145	68.103	68.061	68.020	67.978
.58	67.937	67.895	67.853	67.811	67.770	67.728	67.686	67.644	67.603	67.561
.59	67.519	67.477	67.435	67.393	67.351	67.309	67.267	67.225	67.183	67.141
.60	67.099	67.057	67.014	66.972	66.930	66.888	66.845	66.803	66.761	66.719
.61	66.676	66.634	66.591	66.549	66.506	66.464	66.421	66.379	66.336	66.294
.62	66.251	66.208	66.166	66.123	66.080	66.037	65.994	65.952	65.909	65.866
.63	65.823	65.780	65.737	65.694	65.651	65.608	65.565	65.522	65.479	65.436
.64	65.393	65.349	65.306	65.262	65.219	65.176	65.132	65.089	65.046	65.002
.65	64.959	64.915	64.872	64.828	64.784	64.741	64.697	64.653	64.610	64.566
.66	64.522	64.478	64.434	64.390	64.346	64.302	64.258	64.214	64.170	64.127
.67	64.083	64.038	63.994	63.950	63.905	63.861	63.817	63.772	63.728	63.684
.68	63.639	63.595	63.550	63.506	63.461	63.416	63.372	63.327	63.282	63.238
.69	63.193	63.148	63.103	63.058	63.013	62.968	62.923	62.878	62.833	62.788
.70	62.743	62.698	62.652	62.607	62.562	62.516	62.471	62.426	62.380	62.335
.71	62.289	62.243	62.198	62.152	62.106	62.061	62.015	61.969	61.923	61.877
.72	61.831	61.785	61.739	61.693	61.647	61.601	61.555	61.508	61.462	61.416
.73	61.369	61.323	61.276	61.230	61.183	61.137	61.090	61.043	60.997	60.950
.74	60.903	60.856	60.809	60.762	60.715	60.668	60.621	60.574	60.527	60.479
.75	60.432	60.385	60.337	60.290	60.243	60.195	60.147	60.100	60.052	60.004
.76	59.957	59.909	59.861	59.813	59.765	59.717	59.669	59.621	59.572	59.524
.77	59.476	59.428	59.379	59.331	59.282	59.234	59.185	59.136	59.087	59.039
.78	58.990	58.941	58.892	58.843	58.794	58.745	58.695	58.646	58.597	58.547
.79	58.498	58.449	58.399	58.349	58.300	58.250	58.200	58.150	58.100	58.050
.80	58.000	57.950	57.900	57.850	57.799	57.749	57.699	57.648	57.597	57.547
.81	57.496	57.445	57.394	57.343	57.293	57.242	57.190	57.139	57.088	57.037
.82	56.985	56.934	56.882	56.830	56.779	56.727	56.675	56.623	56.571	56.519
.83	56.467	56.414	56.362	56.310	56.257	56.205	56.152	56.099	56.046	55.994
.84	55.941	55.887	55.834	55.781	55.728	55.675	55.621	55.567	55.514	55.460
.85	55.406	55.352	55.298	55.244	55.190	55.135	55.081	55.026	54.972	54.917
.86	54.862	54.807	54.752	54.697	54.642	54.587	54.531	54.476	54.420	54.364
.87	54.309	54.253	54.196	54.140	54.084	54.028	53.971	53.914	53.858	53.801
.88	53.744	53.687	53.630	53.572	53.515	53.458	53.400	53.342	53.284	53.226
.89	53.168	53.109	53.051	52.992	52.933	52.875	52.816	52.756	52.697	52.638
.90	52.578	52.518	52.458	52.398	52.338	52.278	52.217	52.156	52.096	52.035
.91	51.974	51.912	51.851	51.789	51.727	51.666	51.603	51.541	51.478	51.416
.92	51.353	51.290	51.227	51.163	51.099	51.035	50.971	50.907	50.843	50.778
.93	50.713	50.648	50.582	50.517	50.451	50.385	50.319	50.252	50.185	50.118
.94	50.051	49.983	49.915	49.847	49.779	49.710	49.641	49.572	49.503	49.433
.95	49.362	49.292	49.221	49.150	49.079	49.007	48.934	48.862	48.789	48.715
.96	48.642	48.568	48.493	48.418	48.343	48.267	48.191	48.113	48.036	47.958
.97	47.880	47.801	47.722	47.641	47.561	47.479	47.397	47.314	47.231	47.146
.98	47.062	46.976	46.889	46.801	46.712	46.622	46.531	46.439	46.345	46.250
.99	46.154	46.055	45.955	45.852	45.747	45.639	45.527	(refer to table below)		

.9960	45.527	.9984	45.237	.9992	45.129
.9965	45.470	.9985	45.224	.9993	45.114
.9970	45.411	.9986	45.211	.9994	45.100
.9975	45.351	.9987	45.198	.9995	45.085
.9980	45.289	.9988	45.184	.9996	45.069
.9981	45.276	.9989	45.171	.9997	45.054
.9982	45.263	.9990	45.157	.9998	45.037
.9983	45.250	.9991	45.143	.9999	45.020
				1.0000	45.000

TABLE III—RADIANS PHASE (±.000015) FOR SEMI-INFINITE ATTENUATION SLOPE $k = 1$ $f < f$

f/f_0	0	1	2	3	4	5	6	7	8	9
.00	0.00000	0.00064	0.00127	0.00191	0.00255	0.00318	0.00382	0.00446	0.00509	0.0057
.01	0.00637	0.00700	0.00764	0.00828	0.00891	0.00955	0.01019	0.01082	0.01146	0.0121
.02	0.01273	0.01337	0.01401	0.01464	0.01528	0.01592	0.01655	0.01719	0.01783	0.0184
.03	0.01910	0.01974	0.02037	0.02101	0.02165	0.02228	0.02292	0.02356	0.02419	0.0248
.04	0.02547	0.02611	0.02674	0.02738	0.02802	0.02865	0.02929	0.02993	0.03057	0.0312
.05	0.03184	0.03248	0.03311	0.03375	0.03439	0.03503	0.03566	0.03630	0.03694	0.0375
.06	0.03821	0.03885	0.03949	0.04012	0.04076	0.04140	0.04204	0.04267	0.04331	0.0439
.07	0.04459	0.04523	0.04586	0.04650	0.04714	0.04778	0.04841	0.04905	0.04969	0.0503
.08	0.05097	0.05160	0.05224	0.05288	0.05352	0.05416	0.05479	0.05543	0.05607	0.0567
.09	0.05735	0.05799	0.05862	0.05926	0.05990	0.06054	0.06118	0.06182	0.06245	0.0630
.10	0.06373	0.06437	0.06501	0.06565	0.06629	0.06693	0.06757	0.06821	0.06885	0.0694
.11	0.07013[2]	0.07076	0.07140	0.07204	0.07268	0.07332	0.07396	0.07460	0.07524	0.0758
.12	0.07652	0.07716	0.07780	0.07844	0.07908	0.07972	0.08036	0.08100	0.08164	0.0822
.13	0.08292	0.08356	0.08420	0.08484	0.08548	0.08612	0.08676	0.08740	0.08804	0.0886
.14	0.08932	0.08996	0.09061	0.09125	0.09189	0.09253	0.09317	0.09381	0.09445	0.0950
.15	0.09574[3]	0.09638	0.09702	0.09766	0.09830	0.09894	0.09959	0.10023	0.10087	0.1015
.16	0.10215	0.10279	0.10344	0.10408	0.10472	0.10537	0.10601	0.10665	0.10729	0.1079
.17	0.10858	0.10922	0.10987	0.11051	0.11115	0.11179	0.11244	0.11308	0.11372	0.1143
.18	0.11501	0.11565	0.11630	0.11694	0.11759	0.11823	0.11888	0.11952	0.12016	0.1208
.19	0.12145	0.12210	0.12274	0.12339	0.12403	0.12468	0.12532	0.12596	0.12661	0.1272
.20	0.12790	0.12854	0.12919	0.12984	0.13048	0.13113	0.13178	0.13242	0.13307	0.1337
.21	0.13436	0.13501	0.13565	0.13630	0.13695	0.13759	0.13824	0.13888	0.13953	0.1401
.22	0.14082	0.14147	0.14212	0.14277	0.14342	0.14406	0.14471	0.14536	0.14601	0.1466
.23	0.14730	0.14795	0.14860	0.14925	0.14990	0.15055	0.15119	0.15184	0.15249	0.1531
.24	0.15379	0.15444	0.15509	0.15574	0.15639	0.15704	0.15769	0.15834	0.15899	0.1596
.25	0.16029	0.16094	0.16159	0.16224	0.16289	0.16354	0.16419	0.16484	0.16549	0.1661
.26	0.16680	0.16745	0.16810	0.16875	0.16941	0.17006	0.17071	0.17137	0.17202	0.1726
.27	0.17332	0.17398	0.17463	0.17528	0.17593	0.17659	0.17724	0.17789	0.17855	0.1792
.28	0.17985	0.18051	0.18116	0.18182	0.18247	0.18313	0.18378	0.18444	0.18509	0.1857
.29	0.18641[0]	0.18706	0.18772	0.18837	0.18903	0.18968	0.19034	0.19099	0.19165	0.1923
.30	0.19296	0.19362	0.19428	0.19493	0.19559	0.19625	0.19691	0.19757	0.19823	0.1988
.31	0.19954	0.20020	0.20086	0.20152	0.20218	0.20283	0.20349	0.20415	0.20481	0.2054
.32	0.20613	0.20679	0.20745	0.20811	0.20877	0.20943	0.21009	0.21076	0.21142	0.2120
.33	0.21274[3]	0.21340	0.21406	0.21472	0.21538	0.21605	0.21671	0.21737	0.21803	0.2186
.34	0.21935	0.22002	0.22068	0.22135	0.22201	0.22268	0.22334	0.22401	0.22467	0.2253
.35	0.226[50][0][9]	0.22666	0.22733	0.22799	0.22866	0.22932	0.22999	0.23065	0.23132	0.2319
.36	0.23265	0.23332	0.23398	0.23465	0.23532	0.23599	0.23666	0.23732	0.23799	0.2386
.37	0.23933[2]	0.24000	0.24067	0.24134	0.24200	0.24267	0.24334	0.24401	0.24468	0.2453
.38	0.24601	0.24669	0.24736	0.24803	0.24870	0.24937	0.25005	0.25072	0.25139	0.2520
.39	0.25274[3]	0.25341	0.25408	0.25475	0.25542	0.25610	0.25677	0.25744	0.25811	0.2587
.40	0.25946	0.26013	0.26081	0.26148	0.26216	0.26284	0.26351	0.26419	0.26486	0.2655
.41	0.26621	0.26689	0.26757	0.26824	0.26892	0.26960	0.27028	0.27095	0.27163	0.2723
.42	0.27299	0.27367	0.27435	0.27503	0.27571	0.27639	0.27706	0.27774	0.27842	0.2791
.43	0.27978	0.28047	0.28115	0.28183	0.28251	0.28319	0.28388	0.28456	0.28524	0.2859
.44	0.28660[1]	0.28729	0.28797	0.28866	0.28934	0.29003	0.29071	0.29140	0.29208	0.2927
.45	0.29345	0.29414	0.29482	0.29551	0.29620	0.29688	0.29757	0.29826	0.29894	0.2996
.46	0.30032	0.30101	0.30170	0.30239	0.30308	0.30377	0.30446	0.30515	0.30584	0.3065
.47	0.30721[2]	0.30791	0.30860	0.30929	0.30998	0.31068	0.31137	0.31206	0.31275	0.3134
.48	0.31414	0.31483	0.31553	0.31622	0.31692	0.31761	0.31831	0.31900	0.31970	0.3203
.49	0.32109	0.32179	0.32248	0.32318	0.32388	0.32458	0.32527	0.32597	0.32667	0.3273
.50	0.32807	0.32877	0.32947	0.33017	0.33087	0.33157	0.33227	0.33297	0.33368	0.3343
.51	0.33508	0.33578	0.33648	0.33719	0.33789	0.33860	0.33930	0.34000	0.34071	0.3414
.52	0.34212	0.34282	0.34353	0.34424	0.34495	0.34565	0.34636	0.34707	0.34777	0.3484
.53	0.34919	0.34990	0.35061	0.35132	0.35203	0.35274	0.35345	0.35416	0.35487	0.3555
.54	0.35629	0.35701	0.35772	0.35844	0.35915	0.35986	0.36058	0.36129	0.36201	0.3627

Superscripts—Corrington's values.

TABLE III—*Continued*

f/f_0	0	1	2	3	4	5	6	7	8	9
.55	0.36343	0.36415	0.36487	0.36559	0.36631	0.36702	0.36774	0.36846	0.36918	0.36989
.56	0.37061	0.37133	0.37205	0.37277	0.37350	0.37422	0.37494	0.37566	0.37638	0.37710
.57	0.37782	0.37855	0.37927	0.38000	0.38072	0.38145	0.38217	0.38290	0.38362	0.38435
.58	0.38507	0.38580	0.38653	0.38726	0.38799	0.38872	0.38945	0.39018	0.39091	0.39164
.59	0.39237	0.39310	0.39383	0.39457	0.39530	0.39603	0.39677	0.39750	0.39823	0.39897
.60	0.39970	0.40044	0.40117	0.40191	0.40265	0.40339	0.40413	0.40486	0.40560	0.40634
.61	0.40708	0.40782	0.40856	0.40930	0.41004	0.41079	0.41153	0.41227	0.41301	0.41375
.62	0.41450	0.41524	0.41599	0.41674	0.41748	0.41823	0.41898	0.41972	0.42047	0.42122
.63	0.42197	0.42272	0.42347	0.42422	0.42497	0.42572	0.42647	0.42723	0.42798	0.42873
.64	0.42948	0.43024	0.43100	0.43175	0.43251	0.43327	0.43402	0.43478	0.43554	0.43629
.65	0.43705	0.43781	0.43857	0.43934	0.44010	0.44086	0.44162	0.44238	0.44315	0.44391
.66	0.44467	0.44544	0.44620	0.44697	0.44774	0.44851	0.44927	0.45004	0.45081	0.45158
.67	0.45234	0.45312	0.45389	0.45466	0.45544	0.45621	0.45698	0.45776	0.45853	0.45930
.68	0.46008	0.46086	0.46164	0.46242	0.46319	0.46397	0.46475	0.46553	0.46631	0.46709
.69	0.46787	0.46866	0.46944	0.47023	0.47101	0.47180	0.47258	0.47337	0.47415	0.47494
.70	0.47573	0.47652	0.47731	0.47810	0.47889	0.47968	0.48047	0.48127	0.48206	0.48285
.71	0.48365	0.48444	0.48524	0.48604	0.48684	0.48763	0.48843	0.48923	0.49004	0.49084
.72	0.49164	0.49244	0.49325	0.49405	0.49485	0.49566	0.49647	0.49727	0.49808	0.49889
.73	0.49970	0.50051	0.50132	0.50213	0.50295	0.50376	0.50457	0.50539	0.50621	0.50702
.74	0.50784	0.50866	0.50948	0.51030	0.51112	0.51194	0.51276	0.51358	0.51441	0.51523
.75	0.51605	0.51688	0.51771	0.51854	0.51937	0.52019	0.52103	0.52186	0.52269	0.52352
.76	0.52436	0.52519	0.52603	0.52686	0.52770	0.52854	0.52938	0.53022	0.53106	0.53190
.77	0.53274[5]	0.53359	0.53444	0.53528	0.53613	0.53697	0.53783	0.53868	0.53953	0.54038
.78	0.54123	0.54208	0.54294	0.54380	0.54465	0.54551	0.54637	0.54723	0.54809	0.54895
.79	0.54981	0.55068	0.55154	0.55241	0.55327	0.55414	0.55501	0.55588	0.55676	0.55763
.80	0.55850	0.55938	0.56025	0.56113	0.56201	0.56288	0.56377	0.56465	0.56553	0.56642
.81	0.56730	0.56819	0.56908	0.56996	0.57085	0.57174	0.57264	0.57353	0.57443	0.57532
.82	0.57622	0.57712	0.57802	0.57892	0.57982	0.58072	0.58163	0.58254	0.58345	0.58436
.83	0.58526	0.58618	0.58709	0.58801	0.58892	0.58984	0.59076	0.59168	0.59260	0.59353
.84	0.59445	0.59538	0.59631	0.59723	0.59816	0.59909	0.60003	0.60097	0.60190	0.60284
.85	0.60378	0.60472	0.60567	0.60661	0.60756	0.60850	0.60945	0.61041	0.61136	0.61231
.86	0.61327	0.61423	0.61519	0.61615	0.61711	0.61808	0.61905	0.62002	0.62099	0.62196
.87	0.62293	0.62391	0.62489	0.62587	0.62685	0.62783	0.62882	0.62981	0.63080	0.63179
.88	0.63278	0.63378	0.63478	0.63578	0.63678	0.63779	0.63880	0.63981	0.64082	0.64183
.89	0.64284	0.64387	0.64489	0.64591	0.64693	0.64796	0.64899	0.65003	0.65106	0.65210
.90	0.65313	0.65418	0.65523	0.65628	0.65733	0.65837	0.65943	0.66050	0.66156	0.66262
.91	0.66368	0.66476	0.66583	0.66691	0.66798	0.66906	0.67015	0.67124	0.67233	0.67343
.92	0.67452	0.67562	0.67672	0.67783	0.67894	0.68006	0.68118	0.68230	0.68342	0.68456
.93	0.68569	0.68683	0.68797	0.68911	0.69026	0.69141	0.69257	0.69373	0.69490	0.69607
.94	0.69724	0.69843	0.69961	0.70080	0.70199	0.70319	0.70439	0.70560	0.70681	0.70804
.95	0.70926	0.71049	0.71172	0.71297	0.71421	0.71547	0.71673	0.71800	0.71927	0.72055
.96	0.72183	0.72313	0.72443	0.72574	0.72706	0.72838	0.72971	0.73106	0.73240	0.73377
.97	0.73513	0.73651	0.73790	0.73930	0.74070	0.74213	0.74356	0.74501	0.74646	0.74794
.98	0.74942	0.75092	0.75243	0.75397	0.75552	0.75709[8]	0.75867	0.76028	0.76192	0.76358
.99	0.76527	0.76698	0.76874[3]	0.77053	0.77236	0.77425	0.77620	(refer to table below)		

.9960	0.77620	.9984	0.78125	.9992	0.78315
.9965	0.77720	.9985	0.78148	.9993	0.78340
.9970	0.77822	.9986	0.78171	.9994	0.78366
.9975	0.77928	.9987	0.78195	.9995	0.78392
.9980	0.78036	.9988	0.78218	.9996	0.78419
.9981	0.78058	.9989	0.78242	.9997	0.78446
.9982	0.78080	.9990	0.78266	.9998	0.78475
.9983	0.78103	.9991	0.78290	.9999	0.78505
				1.0000	0.78540

TABLE IV—RADIANS PHASE (±.000015) FOR SEMI-INFINITE ATTENUATION SLOPE $k = 1$
$f > f_0$

f_0/f	0	1	2	3	4	5	6	7	8	9
.00	1.57080	1.57016	1.56952	1.56889	1.56825	1.56761	1.56698	1.56634	1.56570	1.56507
.01	1.56443	1.56379	1.56316	1.56252	1.56188	1.56125	1.56061	1.55997	1.55934	1.55870
.02	1.55806	1.55743	1.55679	1.55615	1.55552	1.55488	1.55424	1.55361	1.55297	1.55233
.03	1.55170	1.55106	1.55042	1.54979	1.54915	1.54851	1.54787	1.54724	1.54660	1.54596
.04	1.54533	1.54469	1.54405	1.54342	1.54278	1.54214	1.54150	1.54087	1.54023	1.53959
.05	1.53896	1.53832	1.53768	1.53704	1.53641	1.53577	1.53513	1.53450	1.53386	1.53322
.06	1.53258	1.53195	1.53131	1.53067	1.53003	1.52940	1.52876	1.52812	1.52748	1.52685
.07	1.52621	1.52557	1.52493	1.52430	1.52366	1.52302	1.52238	1.52174	1.52111	1.52047
.08	1.51983	1.51919	1.51855	1.51792	1.51728	1.51664	1.51600	1.51536	1.51472	1.51409
.09	1.51345	1.51281	1.51217	1.51153	1.51089	1.51026	1.50962	1.50898	1.50834	1.50770
.10	1.50706	1.50642	1.50578	1.50515	1.50451	1.50387	1.50323	1.50259	1.50195	1.50131
.11	1.50067	1.50003	1.49939	1.49875	1.49811	1.49748	1.49684	1.49620	1.49556	1.49492
.12	1.49428	1.49364	1.49300	1.49236	1.49172	1.49108	1.49044	1.48980	1.48916	1.48852
.13	1.48788	1.48724	1.48660	1.48596	1.48532	1.48468	1.48404	1.48339	1.48275	1.48211
.14	1.48147	1.48083	1.48019	1.47955	1.47891	1.47827	1.47763	1.47698	1.47634	1.47570
.15	1.47506	1.47442	1.47378	1.47314	1.47249	1.47185	1.47121	1.47057	1.46993	1.46929
.16	1.46864	1.46800	1.46736	1.46672	1.46607	1.46543	1.46479	1.46414	1.46350	1.46286
.17	1.46222	1.46157	1.46093	1.46029	1.45964	1.45900	1.45836	1.45772	1.45707	1.45643
.18	1.45579	1.45514	1.45450	1.45385	1.45321	1.45257	1.45192	1.45128	1.45063	1.44999
.19	1.44934	1.44870	1.44805	1.44741	1.44677	1.44612	1.44548	1.44483	1.44419	1.44354
.20	1.44290	1.44225	1.44161	1.44096	1.44031	1.43967	1.43902	1.43837	1.43773	1.43708
.21	1.43644	1.43579	1.43514	1.43450	1.43385	1.43320	1.43256	1.43191	1.43127	1.43062
.22	1.42997	1.42933	1.42868	1.42803	1.42738	1.42673	1.42608	1.42544	1.42479	1.42414
.23	1.42349	1.42284	1.42219	1.42155	1.42090	1.42025	1.41960	1.41895	1.41831	1.41766
.24	1.41701	1.41636	1.41571	1.41506	1.41441	1.41376	1.41311	1.41246	1.41181	1.41116
.25	1.41050	1.40985	1.40920	1.40855	1.40790	1.40725	1.40660	1.40595	1.40530	1.40465
.26	1.40400	1.40335	1.40270	1.40204	1.40139	1.40074	1.40008	1.39943	1.39878	1.39813
.27	1.39747	1.39682	1.39617	1.39551	1.39486	1.39421	1.39356	1.39290	1.39225	1.39160
.28	1.39094	1.39029	1.38963	1.38898	1.38832	1.38767	1.38701	1.38636	1.38570	1.38505
.29	1.38439	1.38374	1.38308	1.38242	1.38177	1.38111	1.38046	1.37980	1.37915	1.37849
.30	1.37784	1.37718	1.37652	1.37586	1.37520	1.37455	1.37389	1.37323	1.37257	1.37191
.31	1.37125	1.37060	1.36994	1.36928	1.36862	1.36796	1.36730	1.36665	1.36599	1.36533
.32	1.36467	1.36401	1.36335	1.36269	1.36203	1.36136	1.36070	1.36004	1.35938	1.35872
.33	1.35806	1.35740	1.35673	1.35607	1.35541	1.35475	1.35409	1.35343	1.35277	1.35210
.34	1.35144	1.35078	1.35011	1.34945	1.34878	1.34812	1.34745	1.34679	1.34613	1.34546
.35	1.34480	1.34413	1.34347	1.34280	1.34214	1.34147	1.34081	1.34014	1.33948	1.33881
.36	1.33815	1.33748	1.33681	1.33614	1.33548	1.33481	1.33414	1.33347	1.33280	1.33213
.37	1.33147	1.33080	1.33013	1.32946	1.32879	1.32812	1.32746	1.32679	1.32612	1.32545
.38	1.32478	1.32411	1.32344	1.32277	1.32209	1.32142	1.32075	1.32008	1.31941	1.31873
.39	1.31806	1.31739	1.31672	1.31604	1.31537	1.31470	1.31403	1.31336	1.31268	1.31201
.40	1.31134	1.31066	1.30999	1.30931	1.30864	1.30796	1.30729	1.30661	1.30594	1.30526
.41	1.30458	1.30391	1.30323	1.30255	1.30187	1.30120	1.30052	1.29984	1.29916	1.29849
.42	1.29781	1.29713	1.29645	1.29577	1.29509	1.29441	1.29373	1.29305	1.29237	1.29169
.43	1.29101	1.29033	1.28965	1.28897	1.28828	1.28760	1.28692	1.28624	1.28556	1.28487
.44	1.28419	1.28351	1.28282	1.28214	1.28145	1.28077	1.28008	1.27940	1.27872	1.27803
.45	1.27735	1.27666	1.27597	1.27529	1.27460	1.27391	1.27323	1.27254	1.27185	1.27116
.46	1.27048	1.26979	1.26910	1.26841	1.26772	1.26703	1.26634	1.26565	1.26496	1.26427
.47	1.26358	1.26289	1.26220	1.26150	1.26081	1.26012	1.25943	1.25874	1.25804	1.25735
.48	1.25666	1.25596	1.25527	1.25457	1.25388	1.25318	1.25249	1.25179	1.25110	1.25040
.49	1.24971	1.24901	1.24831	1.24762	1.24692	1.24622	1.24552	1.24482	1.24413	1.24343
.50	1.24273	1.24203	1.24133	1.24063	1.23992	1.23922	1.23852	1.23782	1.23712	1.23642
.51	1.23572	1.23502	1.23431	1.23361	1.23290	1.23220	1.23150	1.23079	1.23009	1.22938
.52	1.22868	1.22797	1.22726	1.22656	1.22585	1.22514	1.22444	1.22373	1.22302	1.22231
.53	1.22161	1.22090	1.22019	1.21948	1.21876	1.21805	1.21734	1.21663	1.21592	1.21521
.54	1.21450	1.21379	1.21307	1.21236	1.21165	1.21093	1.21022	1.20950	1.20879	1.20808

TABLE IV—*Continued*

f_0/f	0	1	2	3	4	5	6	7	8	9
.55	1.20736	1.20664	1.20593	1.20521	1.20449	1.20377	1.20306	1.20234	1.20162	1.20090
.56	1.20019	1.19946	1.19874	1.19802	1.19730	1.19658	1.19586	1.19514	1.19442	1.19369
.57	1.19297	1.19225	1.19152	1.19080	1.19007	1.18935	1.18862	1.18790	1.18717	1.18645
.58	1.18572	1.18499	1.18426	1.18353	1.18280	1.18208	1.18135	1.18062	1.17989	1.17916
.59	1.17843	1.17770	1.17696	1.17623	1.17550	1.17476	1.17403	1.17330	1.17256	1.17183
.60	1.17110	1.17036	1.16962	1.16888	1.16815	1.16741	1.16667	1.16593	1.16519	1.16446
.61	1.16372	1.16298	1.16224	1.16149	1.16075	1.16001	1.15927	1.15853	1.15778	1.15704
.62	1.15630	1.15555	1.15481	1.15406	1.15331	1.15257	1.15182	1.15107	1.15032	1.14958
.63	1.14883	1.14808	1.14733	1.14658	1.14582	1.14507	1.14432	1.14357	1.14282	1.14207
.64	1.14131	1.14056	1.13980	1.13904	1.13829	1.13753	1.13677	1.13602	1.13526	1.13450
.65	1.13375	1.13298	1.13222	1.13146	1.13070	1.12994	1.12917	1.12841	1.12765	1.12689
.66	1.12613	1.12536	1.12459	1.12382	1.12306	1.12229	1.12152	1.12075	1.11999	1.11922
.67	1.11845	1.11768	1.11690	1.11613	1.11536	1.11459	1.11381	1.11304	1.11227	1.11149
.68	1.11072	1.10994	1.10916	1.10838	1.10760	1.10682	1.10604	1.10526	1.10448	1.10371
.69	1.10293	1.10214	1.10135	1.10057	1.09978	1.09900	1.09821	1.09743	1.09664	1.09586
.70	1.09507	1.09428	1.09349	1.09270	1.09191	1.09112	1.09032	1.08953	1.08874	1.08794
.71	1.08715	1.08635	1.08556	1.08476	1.08396	1.08316	1.08236	1.08156	1.08076	1.07996
.72	1.07916	1.07836	1.07755	1.07675	1.07594	1.07514	1.07433	1.07352	1.07271	1.07191
.73	1.07110	1.07029	1.06947	1.06866	1.06785	1.06704	1.06622	1.06541	1.06459	1.06378
.74	1.06296	1.06214	1.06132	1.06050	1.05968	1.05886	1.05804	1.05721	1.05639	1.05557
.75	1.05474	1.05391	1.05309	1.05226	1.05143	1.05060	1.04977	1.04894	1.04811	1.04727
.76	1.04644	1.04560	1.04477	1.04393	1.04309	1.04226	1.04142	1.04058	1.03973	1.03889
.77	1.03805	1.03721	1.03636	1.03551	1.03467	1.03382	1.03297	1.03212	1.03127	1.03042
.78	1.02957	1.02871	1.02786	1.02700	1.02615	1.02529	1.02443	1.02357	1.02271	1.02185
.79	1.02099	1.02012	1.01925	1.01839	1.01752	1.01666	1.01578	1.01491	1.01404	1.01317
.80	1.01230	1.01142	1.01054	1.00967	1.00879	1.00791	1.00703	1.00615	1.00526	1.00438
.81	1.00350	1.00261	1.00172	1.00083	.99994	.99905	.99816	.99726	.99637	.99547
.82	.99458	.99368	.99278	.99188	.99097	.99007	.98916	.98826	.98735	.98644
.83	.98553	.98462	.98370	.98279	.98187	.98096	.98004	.97912	.97819	.97727
.84	.97635	.97542	.97449	.97356	.97263	.97170	.97077	.96983	.96889	.96796
.85	.96702	.96607	.96513	.96418	.96324	.96229	.96134	.96039	.95944	.95848
.86	.95753	.95657	.95561	.95464	.95368	.95272	.95175	.95078	.94981	.94884
.87	.94787	.94689	.94591	.94493	.94395	.94297	.94198	.94099	.93999	.93900
.88	.93801	.93701	.93601	.93501	.93401	.93301	.93200	.93099	.92998	.92896
.89	.92795	.92693	.92591	.92488	.92386	.92284	.92180	.92077	.91973	.91870
.90	.91766	.91662	.91557	.91452	.91347	.91242	.91136	.91030	.90924	.90818
.91	.90712	.90604	.90496	.90389	.90281	.90174	.90064	.89955	.89846	.89737
.92	.89628	.89518	.89407	.89296	.89185	.89074	.88962	.88850	.88737	.88624
.93	.88511	.88397	.88283	.88168	.88054	.87938	.87823	.87706	.87590	.87473
.94	.87355	.87237	.87119	.87000	.86881	.86761	.86641	.86519	.86398	.86276
.95	.86154	.86031	.85907	.85783	.85658	.85533	.85407	.85280	.85153	.85025
.96	.84896	.84766	.84637	.84505	.84374	.84241	.84108	.83974	.83839	.83703
.97	.83567	.83428	.83290	.83150	.83009	.82867	.82724	.82579	.82434	.82286
.98	.82138	.81988	.81836	.81683	.81528	.81371	.81212	.81051	.80888	.80722
.99	.80553	.80381	.80206	.80027	.79844	.79655	.79460	(refer to table below)		

.9960	0.79460	.9984	0.78954	.9992	0.78765
.9965	0.79360	.9985	0.78931	.9993	0.78739
.9970	0.79257	.9986	0.78908	.9994	0.78714
.9975	0.79152	.9987	0.78885	.9995	0.78688
.9980	0.79044	.9988	0.78862	.9996	0.78661
.9981	0.79022	.9989	0.78838	.9997	0.78633
.9982	0.78999	.9990	0.78814	.9998	0.78605
.9983	0.78977	.9991	0.78789	.9999	0.78575
				1.0000	0.78540

of phase versus frequency and drawing a smooth curve weighting the points in accordance with the errors known by experience to occur for various types

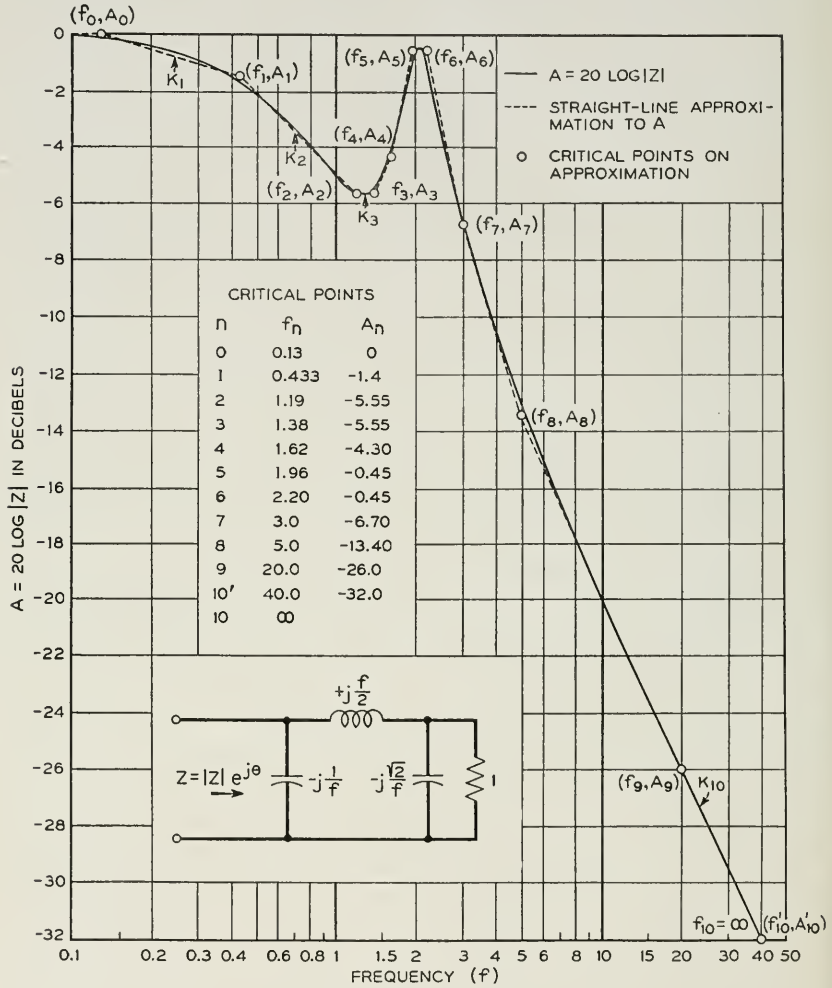


Fig. 5—20 log | Z |.

of departures of the straight line approximation from the exact characteristic.

Although the degree and *db* relationship is applicable to attenuation and phase computations, nepers and radians are proper theoretical units which can be used in other problems[9]. For instance, Tables III and IV give the

[9] Bode, "Network Analysis and Feedback Amplifier Design," Chapter XV, page 340.

reactance in ohms associated with a semi-infinite unit slope of resistance where a unit slope of resistance is one in which a one-ohm change in resistance

TABLE V

TABULATION OF CRITICAL POINTS AND DETERMINATION OF SLOPES OF STRAIGHT LINES APPROXIMATING CHARACTERISTIC OF FIG. 5

n	f_n	A_n	$A_n - A_{n-1}$	$20 \log \frac{f_n}{f_{n-1}}$	k_n
0	.13	0	—	—	—
1	.433	−1.40	−1.40	10.45	−.134
2	1.19	−5.55	−4.15	8.78	−.473
3	1.38	−5.55	.00	1.287	0
4	1.62	−4.30	+1.25	1.393	+.897
5	1.96	−.45	+3.85	1.655	+2.326
6	2.20	−.45	.00	1.003	0
7	3.00	−6.70	−6.25	2.694	−2.320
8	5.00	−13.40	−6.70	4.437	−1.510
9	20.0	−26.00	−12.60	12.04	−1.046
10′	40.0	−32.0	−6.0	6.02	
10	∞				−1.0

Note that $f'_{10} = 40.0$ is chosen to get k_{10} over a finite section of the semi-infinite slope extending to $f = \infty$.

TABLE VI

SUMMATION OF PHASE ASSOCIATED WITH 20 LOG $|Z|$ OF FIG. 5 AT $f = 1.5$

n	f_n from Table V	$\frac{f_n}{f}$	$\frac{f}{f_n}$	θ_n Degrees	$\theta_{n-1} - \theta_n$ Degrees	k_n from Table V	$k_n(\theta_{n-1} - \theta_n)$ Degrees
0	.13	.087		86.824			
1	.433	.289		79.357	7.467	−.134	−1.00
2	1.19	.793		58.349	21.008	−.473	−9.94
3	1.38	.920		51.353	6.996	0	
4	1.62		.926	39.029	12.324	+.897	+11.05
5	1.96		.765	30.283	8.746	+2.326	+20.34
6	2.20		.682	26.450	3.833	0	
7	3.00		.5	18.797	7.653	−2.320	−17.75
8	5.00		.3	11.056	7.741	−1.510	−11.69
9	20.0		.075	2.737	8.319	−1.046	−8.70
10	∞		0	.000	2.737	−1.00	−2.74
						$\Sigma k_n (\theta_{n-1} - \theta_n)$	= −20.43
						$\theta (f = 1.5)$	= −20.4°

Note that for f_0 to f_3 the ratio of f to f_n must be taken f_n/f to be less than unity and θ_n is therefore read from Table II, whereas for f_4 to f_{10} the ratio must be taken f/f_n and θ_n is therefore read from Table I.

occurs between frequencies which are in the ratio $e = 2.7183$. The same technique described above for the determination of the phase associated

TABLE VII

CRITICAL POINTS FOR FIVE LINE APPROXIMATION TO CHARACTERISTIC OF FIG. 5

n	f_n	A_n
0	.25	0
1	1.40	−5.8
2	2.10	0
3	3.00	−7.0
4	10.0	−20.0
5′	40.0	−32.0
5	∞	

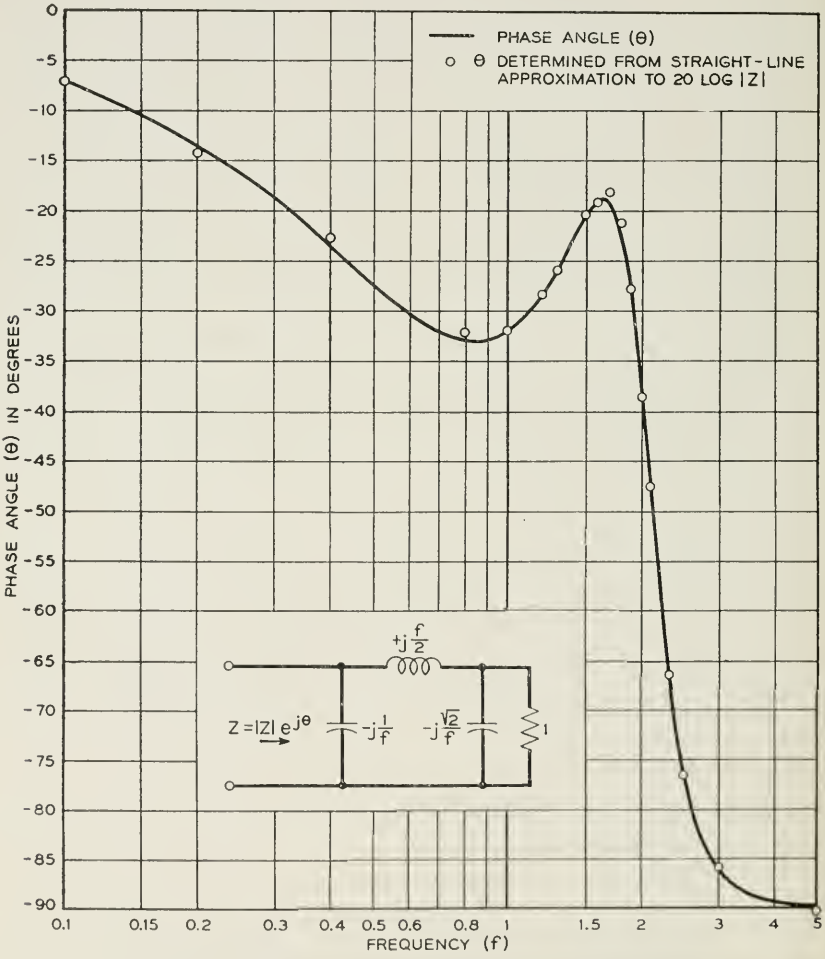


Fig. 6—Phase associated with 20 log | Z | of Fig. 5.

with a given attenuation characteristic may therefore be used to determine the reactance associated with a given resistance characteristic. The only

TABLE VIII
TABULATION OF CRITICAL POINTS AND DETERMINATION OF SLOPES OF STRAIGHT LINES APPROXIMATING RESISTANCE CHARACTERISTIC OF FIG. 7

n	f_n	R_n	$R_n - R_{n-1}$	$2.303 \log \frac{f_n}{f_{n-1}}$	k_n
0	.078	1.000	0		
1	.185	.912	−.088	.864	−.102
2	.290	.805	−.107	.450	−.238
3	.900	.400	−.405	1.133	−.357
4	1.20	.400	0	—	
5	1.50	.547	+.147	.2231	+.659
6	1.67	.840	+.293	.1074	+2.728
7	1.84	1.280	+.440	.0969	+4.54
8	1.92	1.280	0		
9	2.20	.335	−.945	.1361	−6.94
10	2.45	.094	−.241	.1076	−2.24
11	2.85	.015	−.079	.1512	−.52
12	5.00	.000	−.015	.562	−.027

TABLE IX
SUMMATION OF REACTANCE ASSOCIATED WITH RESISTANCE OF FIG. 7 AT $f = 1.0$

n	f_n (From Table VIII)	$\frac{f_n}{f}$	$\frac{f}{f_n}$	X_n Ohms	$X_{n-1} - X_n$ Ohms	k_n (From Table VIII)	$k_n (X_{n-1} - X_n)$ Ohms
0	.078	.078		1.52111			
1	.185	.185		1.45257	.06854	−.102	−.0070
2	.290	.290		1.38439	.06818	−.238	−.0162
3	.90	.900		.91766	.46673	−.357	−.1666
4	1.20		.833	.58801	.32965		
5	1.50		.667	.45004	.13797	+.659	+.0909
6	1.67		.599	.39897	.05107	+2.728	+.1393
7	1.84		.543	.35844	.04053	+4.54	+.1840
8	1.92		.521	.34282	.01562		
9	2.20		.455	.29688	.04594	−6.94	−.3188
10	2.45		.408	.26486	.03202	−2.24	−.0717
11	2.85		.351	.22666	.03820	−.52	−.0199
12	5.00		.200	.12790	.09876	−.027	−.0027
					$\Sigma k_n (X_{n-1} - X_n) = -.1887$		
					$X (f = 1.0) = -.189$ Ohm		

difference is that the slopes of the straight lines approximating the resistance plotted on a log frequency scale are determined by the expression below:

$$k_n = \frac{R_n - R_{n-1}}{\log_e \frac{f_n}{f_{n-1}}} = \frac{R_n - R_{n-1}}{2.303 \log \frac{f_n}{f_{n-1}}}$$

where:

R_n is the resistance at f_n on the straight line approximation to R.

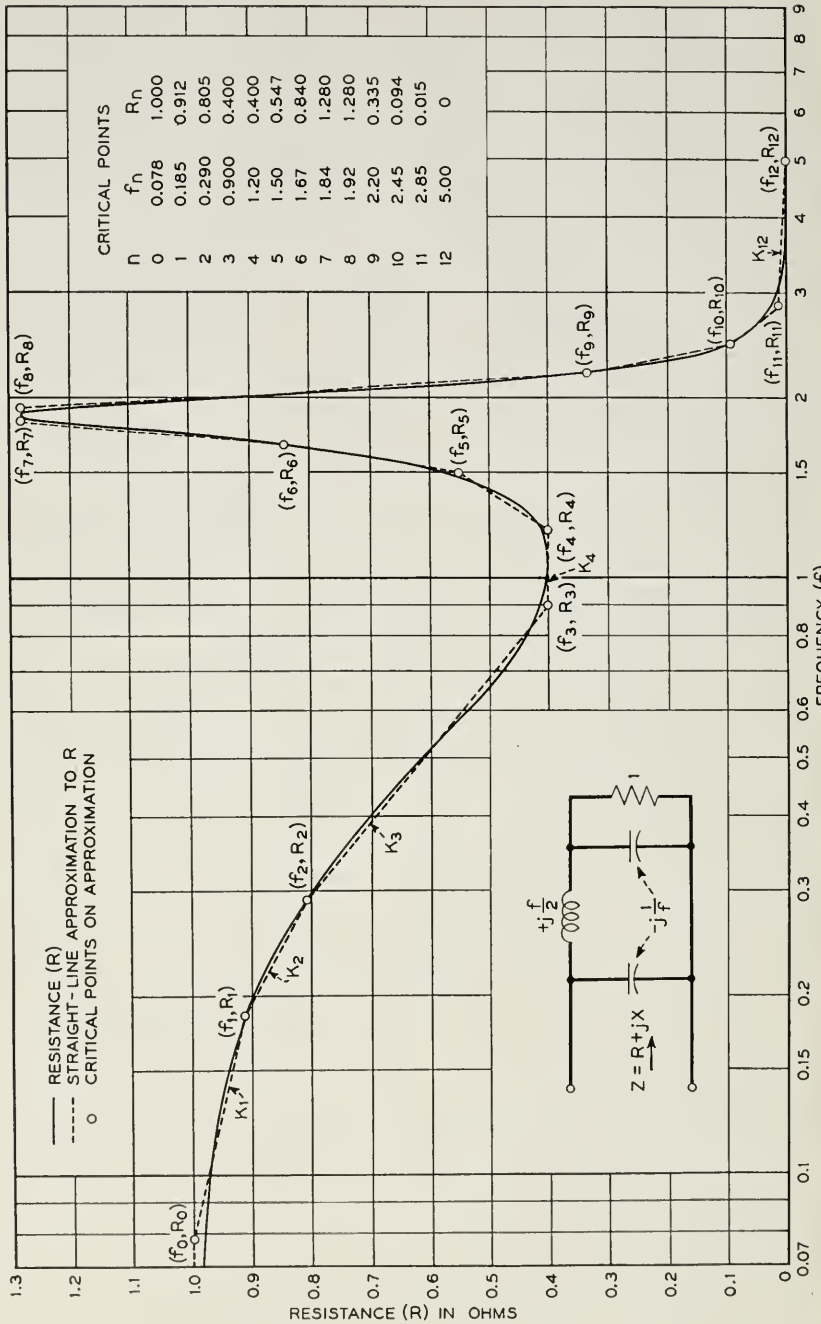


Fig. 7—Resistance characteristic.

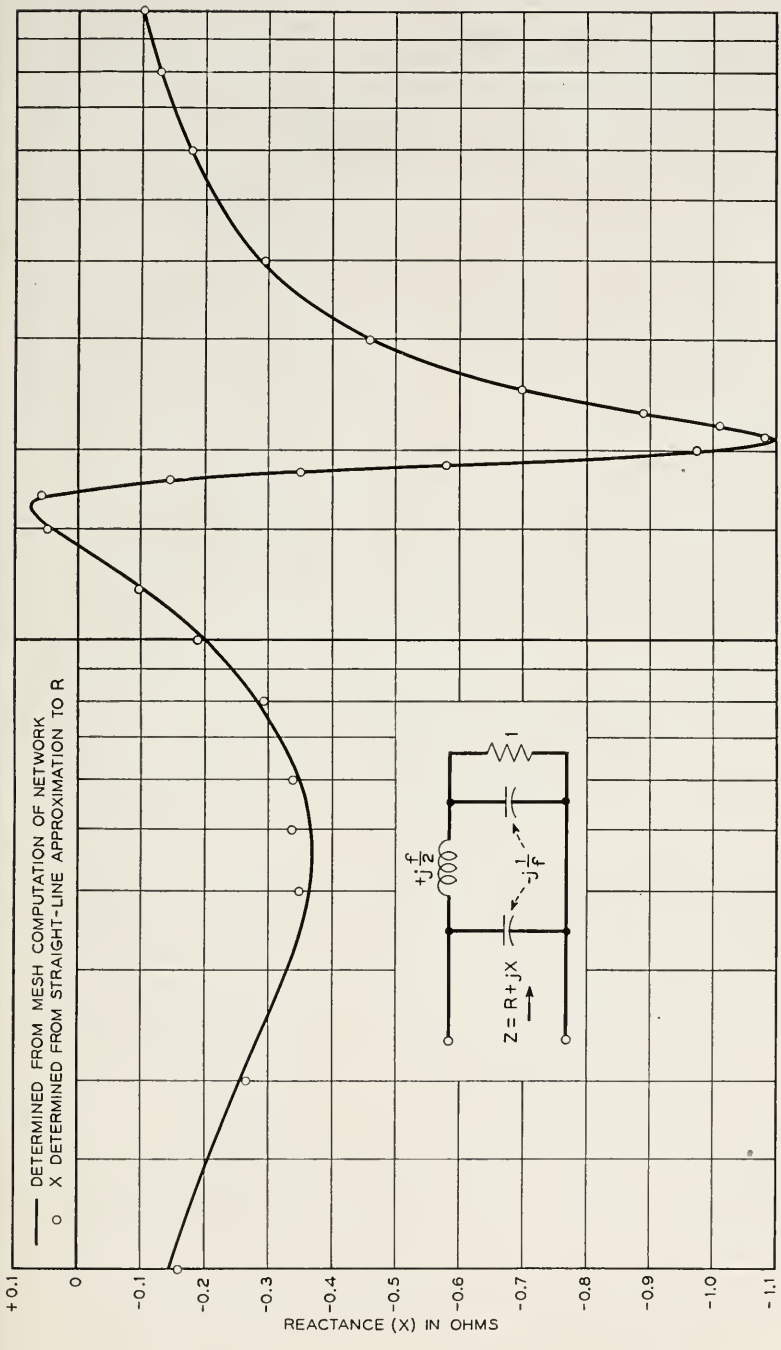


Fig. 8—Reactance associated with resistance characteristic of Fig. 7.

As an example of the determination of the reactance associated with a given resistance characteristic, consider the resistance characteristic of Fig. 7 and the straight line approximation shown in dotted form. The slopes of the straight lines are determined as illustrated in Table VIII.
Having determined the slopes of the various straight lines of the approximation, the reactance can be summed at any desired frequency. As an illustration the reactance is summed at $f = 1.0$, in Table IX.

The mesh computed reactance of the network of Fig. 7 is plotted in Fig. 8 and the reactance summed for $f = 1.0$ is seen to be within .01 ohm of the true reactance. The reactance was summed at a considerable number of frequencies and the results plotted as individual points in Fig. 8. The degree of approximation to the true reactance should be similar to the

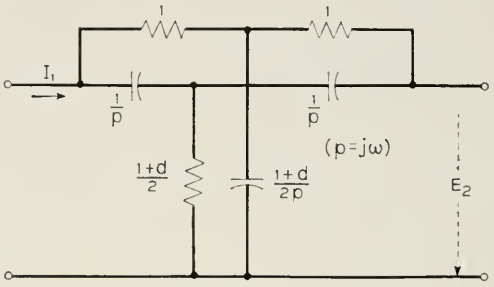


Fig. 9—Parallel T network.

degree of approximation to the original resistance and this is borne out by the example where the straight line approximation to the resistance characteristic is within $\pm$ .03 ohm and the maximum departure of the reactance determined from the straight line approximation is $\pm$ .025 ohm.

As was pointed out in the attenuation example a much simpler straight line approximation to the resistance characteristic would have resulted in a reactance determination without too much greater error than the determination of the illustration.

A word of caution is necessary in connection with the use of the straight line approximation method discussed above. The true phase or reactance is reliably obtained only in those cases where the problem in question is a minimum phase one. In order to illustrate the failure of the method in those problems in which non-minimum phase conditions exist consider the parallel T network of Fig. 9. The transfer impedance Z_{012} defined by the

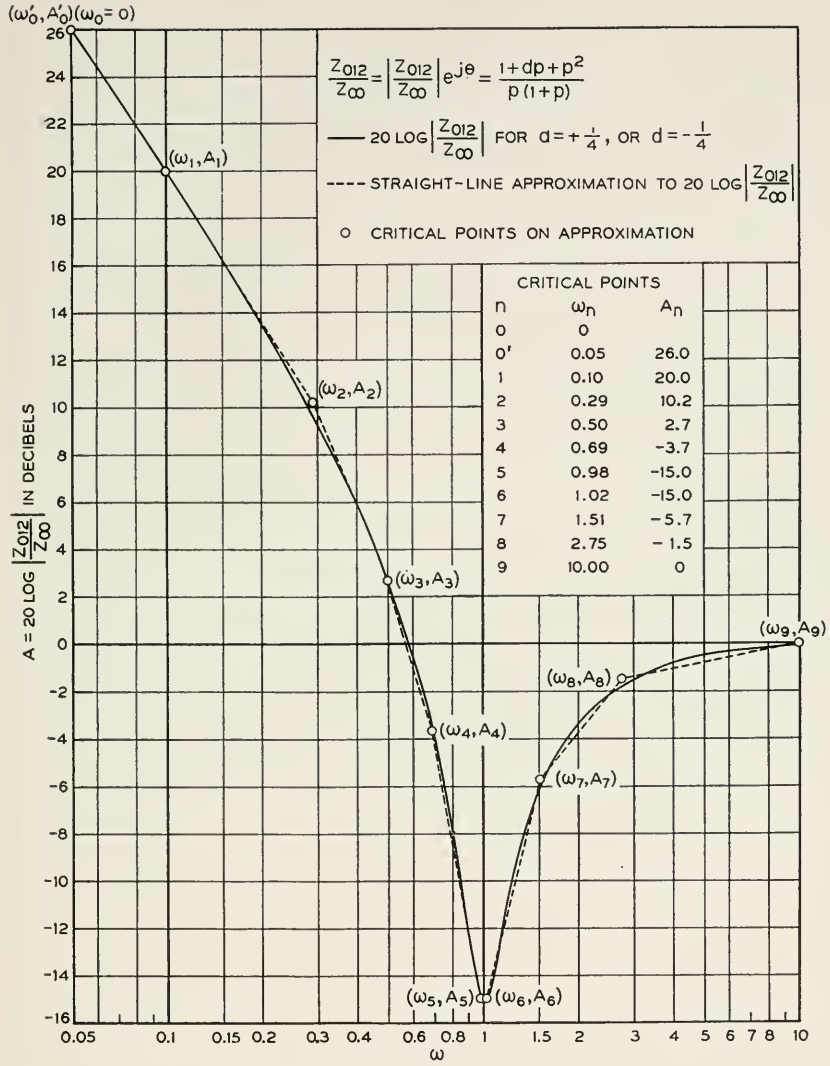


Fig. 10—20 log $\left|\frac{Z_{012}}{Z_\infty}\right|$ for network of Fig. 9.

ratio of the open circuit voltage E_2 to the open circuit driving current I_1 is given by:

$$Z_{012} = \frac{1}{2}\frac{1+d}{2+d}\frac{1+dp+p^2}{p(1+p)}.$$

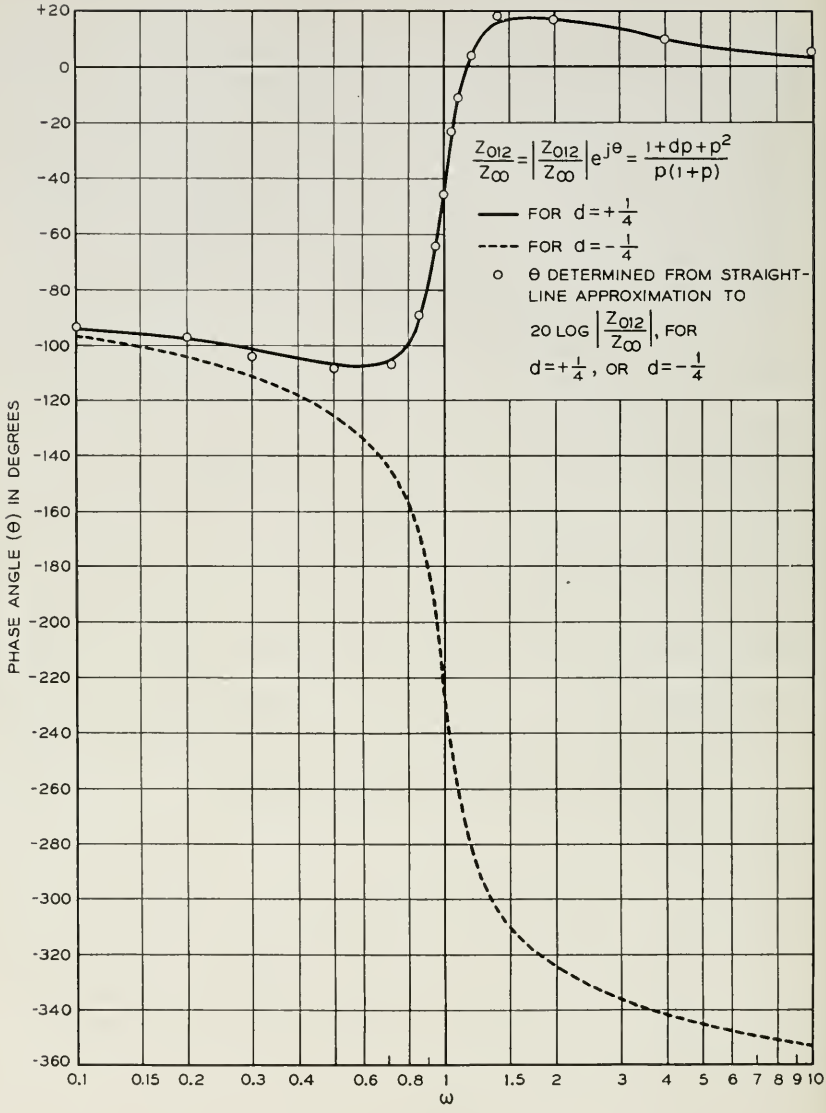


Fig. 11—Phase angle of $\frac{Z_{012}}{Z_\infty}$ for network of Fig. 9.

If we take the ratio of Z_{012} to its value for $\omega = \infty$ then:

$$\frac{Z_{012}}{Z_\infty} = \left| \frac{Z_{012}}{Z_\infty} \right| e^{j\theta} = \frac{1 + dp + p^2}{p(1 + p)} .$$

$20 \log \left| \frac{Z_{012}}{Z_\infty} \right|$ is plotted in Fig. 10 for $d = +1/4$ and it is apparent than $20 \log \left| \frac{Z_{012}}{Z_\infty} \right|$ for $d = -1/4$ is identical. This identity does not hold for θ, however. This is shown in Fig. 11 where θ for $d = +1/4$ and θ for $d = -1/4$ are plotted.

The real characteristic of Fig. 10 was then approximated by a series of straight lines determined by the critical points listed and the phase associated with this straight line approximation summed. The phase so determined is plotted as individual points in Fig. 11. It is seen that this summation determined the phase of the function in question for $d = +1/4$ but completely failed to do so for $d = -1/4$. The function for $d = -1/4$ is an example of a non-minimum phase function for which the above technique fails to determine the phase of the function from its attenuation characteristic.[10]

There are certain instances where the above technique can be usefully applied in connection with non-minimum phase systems in spite of the failure of the method to predict the total phase.[11] However, the necessity of checking for non-minimum phase conditions and, if such exist, determining whether the above method of computing phase is at all applicable, is illustrated by the non-minimum phase example above.

[10] This is the anticipated result since the function is identified as a non-minimum phase function by the fact that it has two zeros falling in the right half p plane.

[11] Bode, "Network Analysis and Feedback Amplifier Design," Chap. XIV, page 309.

Abstracts of Technical Articles by Bell System Authors

Television Network Facilities.[1] L. G. Abraham and H. I. Romnes. Television networks, like sound broadcasting networks, must be available to make distribution of high quality programs economical. For television circuits interconnecting studios in different cities, coaxial cable and radio relay are the most suitable methods. For short distance transmission balanced wire pairs also may be used. Local conditions will control the type circuit selected.

Protective Coatings on Bell System Cables.[2] V. J. Albano and Robert Pope. The practice of placing some Bell System cables directly in the ground without the use of conduit was introduced in about 1929. Since bare cable thus installed would be subject to the corrosive action of soils, or damage from lightning or gophers, suitable protective coatings to guard against these hazards had to be developed. Seven types of such coverings are described, and their particular field of application is indicated.

Surface States and Rectification at a Metal Semi-Conductor Contact.[3] John Bardeen. Localized states (Tamm levels), having energies distributed in the "forbidden" range between the filled band and the conduction band, may exist at the surface of a semi-conductor. A condition of no net charge on the surface atoms may correspond to a partial filling of these states. If the density of surface levels is sufficiently high, there will be an appreciable double layer at the free surface of a semi-conductor formed from a net charge from electrons in surface states and a space charge of opposite sign, similar to that at a rectifying junction, extending into the semi-conductor. This double layer tends to make the work function independent of the height of the Fermi level in the interior (which in turn depends on impurity content). If contact is made with a metal, the difference in work function between metal and semi-conductor is compensated by surface states charge, rather than by a space charge as is ordinarily assumed, so that the space charge layer is independent of the metal. Rectification characteristics are then independent of the metal. These ideas are used to explain results of Meyerhof and others on the relation between contact potential differences and rectification.

[1] *Electrical Engineering*, May 1947.
[2] *Corrosion*, May 1947.
[3] *Phys. Rev.*, May 15, 1947.

Plating on Aluminum.[4] R. A. EHRHARDT* and J. M. GUTHRIE. This article describes tests made to develop a satisfactory process for producing adherent electrodeposits on aluminum alloys using a zincate immersion pretreatment.

Since the major interest was the fabrication of aluminum structures by the use of lead-tin solders the adherence of the deposit was determined by measuring the strength of soldered joints.

Excellent results were obtained with commercially pure aluminum and copper bearing alloys and satisfactory results with magnesium and silicon bearing alloys.

Corrective Networks.[5] F. L. HOPPER. A type of fully compensated constant resistance network is described which provides a larger family of equalization characteristics particularly suited to corrective use in rerecording as determined by aural monitoring.

Spectrochemical Analysis of Ceramics and Other Non-Metallic Materials.[6] EDWIN K. JAYCOX. The procedure described is applicable to the quantitative spectrochemical analysis of ceramics, ashes, ores, paints, and other non-metallic materials for the determination of most of the common metals and their oxides. These include: aluminum, boron, barium, beryllium, calcium, copper, chromium, iron, lithium, magnesium, manganese, sodium, lead, silicon, titanium, zinc, and zirconium, in the general range of 0.30–70.0 per cent. Samples in the form of a fine powder are mixed one part of sample to 10–100 parts of a suitable metal oxide which serves as a buffer, diluent, and internal control. Carbon dust is added to this mixture for its additional buffering effect. Spectra are obtained of the samples and of an appropriate series of standards. Determinations of the amount of element sought are made, in most cases by the well known internal standard technique, in others by the simple comparison standard procedure.

The Spectrochemical Analysis of Nickel Alloys.[7] EDWIN K. JAYCOX. A procedure is described for the analysis of nickel alloys for copper, iron, lead, magnesium, manganese, silicon, titanium, and zinc in the range 0.005–0.30 per cent and for boron in the range 0.0003–0.03 per cent. Samples are taken into solution with dilute nitric acid, evaporated to dryness, and baked at 400°C. The resulting dry nitrate-oxide powder is mixed with pure carbon dust which acts as a buffer and diluent. Aliquots of each sample and of a

[4] *The Monthly Review, American Electroplaters Society*, April 1947.
* Of Bell Tel. Labs.
[5] *Jour. Soc. Motion Pic. Engrs.*, March 1947.
[6] *Jour. Optical Soc. Amer.*, March 1947.
[7] *Jour. Optical Soc. Amer.*, March 1947.

series of standards are excited in the direct current arc, and their spectra recorded on the same plate. Determinations of the amounts of constituent elements present in the sample are made by measuring the logarithm of the ratio of the relative intensities of a line of the element sought to that of a nickel control line by the general internal control technique.

Measurement of the Viscosity and Shear Elasticity of Liquids by Means of a Torsionally Vibrating Crystal.[8] W. P. Mason. This paper describes a method of measuring viscosities of liquids at high frequencies by means of oscillating cylinders, in which a torsionally vibrating crystal generates a viscous wave in the medium to be measured. Both a reactance and a resistance loading occur in the crystal which lowers its frequency and raises the measured resistance at resonance. The viscosity may then be determined by measuring the changes in the properties of the crystal. By varying the voltage on the crystal, the shearing displacement can be varied and hence the viscosity can be measured as a function of shearing stress. Measurements on light oils over a viscosity range from 0.01 poise to 10 poises check within a few per cent when made with rough temperature-control conditions.

Considerations in the Design of Centimeter-Wave Radar Receivers.[9] Stewart E. Miller. A review of the radar duplexer and receiver, as developed during the war, is presented. Attention is devoted to the principles of operation and typical circuit arrangements employed in the duplexer, the crystal converter, the local-oscillator injection circuits, the intermediate-frequency amplifier, and the automatic-tuning unit. Emphasis is placed on methods found advantageous in the 1-centimeter and 3-centimeter wavelength regions. The interrelation between the various receiver components in determining the over-all receiver noise figure is shown analytically, and typical performance numbers are given.

Experimental Rural Radiotelephony.[10] J. Harold Moore, Paul K. Seyler and S. B. Wright. The first rural party-line telephone service utilizing radio installations operating on the subscribers' premises was undertaken experimentally in the vicinity of Cheyenne Wells, near the eastern border of Colorado. Radio links have been used to supply regular telephone service to eight ranches since August 20, 1946. The development of a standard rural radiotelephone system will be aided materially by the experience gained from these experiments.

[8] *Transactions A.S.M.E.*, May 1947.
[9] *Proc. I.R.E.*, April 1947.
[10] *Electrical Engineering*, April 1947.

Alkaline Earth Porcelains Possessing Low Dielectric Loss.[11] M. D. RIGTERINK and R. O. GRISDALE. Alkaline earth porcelains have been prepared from mixtures of clay, flint, and synthetic fluxes consisting of clay calcined with at least three alkaline earth oxides. These porcelains possess excellent dielectric properties, have low coefficients of thermal expansion, are white, and are especially valuable as bases for deposited carbon resistors for which they were developed. Their characteristics make it probable that other uses will be found for materials of this type.

An illustrative composition is 50.0% Florida kaolin, 15.0% flint (325 mesh), 35.0% calcine (200 mesh). The composition of the calcine is 40.0% Florida kaolin, 15.0% $MgCO_3$, 15.0% $CaCO_3$, 15.0% $SrCO_3$, 15.0% $BaCO_3$, calcined at 1200°C. The electrical properties of this body at 1 mc. are Q at 25°C., 2160; Q at 250°C., 280; Q at 350°C., 90; specific resistance at 150°C., $10^{13.5}$ ohm-cm. and at 300°C., $10^{10.7}$ ohm-cm.

Attenuation of Drainage Effects on a Long Uniform Structure with Distributed Drainage.[12] J. M. STANDRING, JR. This paper discusses the general behavior of forced drainage currents on long uniform underground communication cables with particular regard to the case where drainage is applied at regular intervals. Expressions are developed for the structure-to-earth potential which is caused by uniformly spaced drainers when the power supply is from variable e.m.f. sources, such as rectifiers, and also for the case where fixed e.m.f.'s, such as galvanic anodes, are employed.

[11] *Jour. Amer. Ceramic Society*, March 1, 1947.
[12] *Corrosion*, June 1947.

Contributors to this Issue

Clifford E. Fay, B.S. in Electrical Engineering, Washington University, 1925; M.S., 1927. Bell Telephone Laboratories, 1927–. Mr. Fay has been engaged principally in the development of power vacuum tubes for radio purposes.

Laurence W. Morrison, B.S. in Electrical Engineering, University of Wisconsin, 1930. Graduate work, 1930–31. Bell Telephone Laboratories, 1931–. Mr. Morrison was engaged in the development of telephone and television terminal equipment for the coaxial system to 1941. During the past war he was concerned with the development of various radar systems as project engineer. Since 1945 he has been in charge of a group concerned with the development of television transmission over wire facilities.

G. E. Mueller, B.S., Missouri School of Mines and Metallurgy, 1939; M.S., Purdue University, 1940. Bell Telephone Laboratories, 1940–46. Mr. Mueller was engaged in television and radio research. During the war he worked on radar antenna development. Mr. Mueller is now Assistant Professor of Electrical Engineering at the Ohio State University.

Sloan D. Robertson, B.E.E., University of Dayton, 1936; M.Sc., Ohio State University, 1938; Ph.D., 1941, Instructor of Electrical Engineering, University of Dayton, 1940. Bell Telephone Laboratories, 1940–. Dr. Robertson was engaged in microwave radar work in the Radio Research Department during the war. He is now engaged in fundamental microwave radio research.

D. E. Thomas, B.S. in Electrical Engineering, Pennsylvania State College, 1929; M.A., Columbia University, 1932. Bell Telephone Laboratories, 1929–. On Military leave from 1942–46 with U. S. Army Signal Corps and U. S. Army Air Forces. Mr. Thomas has been engaged in investigations of submarine telephone cable systems.

W. A. Tyrrell, B.S., Yale University, 1935; Ph.D., 1939. Bell Telephone Laboratories, 1939–. Dr. Tyrrell has been engaged in waveguide research, principally in the field of microwave loss measurements. During the war he developed a number of waveguide components for Navy radar.